

ORIGINAL ARTICLE

Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information

S. Kim^{1†}, H. S. Kim^{2,3†}, E. Kim¹, M. G. Lee², E.-C. Shin⁴, S. Paik^{1,3} & S. Kim^{1*}

¹Severance Biomedical Science Institute; ²Department of Pharmacology, Pharmacogenomic Research Center for Membrane Transporters, Brain Korea 21 PLUS Project for Medical Sciences; ³Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul; ⁴Graduate School of Medical Science and Engineering, KAIST, Daejeon, Korea

*Correspondence to: Prof. Sangwoo Kim, Severance Biomedical Science Institute, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seoul 120-752, Korea.
Tel: +82-2-2228-0913; E-mail: swkim@yuhs.ac

[†]Both authors contributed equally as co-first authors.

Background: Tumor-specific mutations form novel immunogenic peptides called neoantigens. Neoantigens can be used as a biomarker predicting patient response to cancer immunotherapy. Although a predicted binding affinity (IC₅₀) between peptide and major histocompatibility complex class I is currently used for neoantigen prediction, large number of false-positives exist.

Materials and methods: We developed Neopepsee, a machine-learning-based neoantigen prediction program for next-generation sequencing data. With raw RNA-seq data and a list of somatic mutations, Neopepsee automatically extracts mutated peptide sequences and gene expression levels. We tested 14 immunogenicity features to construct a machine-learning classifier and compared with the conventional methods based on IC₅₀ regarding sensitivity and specificity. We tested Neopepsee on independent datasets from melanoma, leukemia, and stomach cancer.

Results: Nine of the 14 immunogenicity features that are informative and inter-independent were used to construct the machine-learning classifiers. Neopepsee provides a rich annotation of candidate peptides with 87 immunogenicity-related values, including IC₅₀, expression levels of neopeptides and immune regulatory genes (e.g. PD1, PD-L1), matched epitope sequences, and a three-level (high, medium, and low) call for neoantigen probability. Compared with the conventional methods, the performance was improved in sensitivity and especially two- to threefold in the specificity. Tests with validated datasets and independently proven neoantigens confirmed the improved performance in melanoma and chronic lymphocytic leukemia. Additionally, we found sequence similarity in proteins to known pathogenic epitopes to be a novel feature in classification. Application of Neopepsee to 224 public stomach adenocarcinoma datasets predicted ~7 neoantigens per patient, the burden of which was correlated with patient prognosis.

Conclusions: Neopepsee can detect neoantigen candidates with less false positives and be used to determine the prognosis of the patient. We expect that retrieval of neoantigen sequences with Neopepsee will help advance research on next-generation cancer immunotherapies, predictive biomarkers, and personalized cancer vaccines.

Key words: cancer, classification, immunoinformatics, neoantigen, next-generation sequencing

Introduction

Somatic mutations can cause tumor-specific neopeptide fragments (so-called ‘neoantigens’), some of which induce cytotoxic T-cell responses [1]. Importantly, neoantigen prediction can be exploited to identify responders to immune checkpoint inhibitors [2–4]. Therefore, systematic evaluation of somatic mutations may help

advance the promising clinical outcomes of immunotherapies in cancer treatment.

For a somatic mutation to be recognized by cytotoxic T cells, the mutant peptide should be presented by major histocompatibility complex class I (MHC-I) molecules [5] (Figure 1A). First, mutant proteins are cut into short peptides by proteasomes and then

transported into the endoplasmic reticulum by the transporter associated with antigen processing (TAP). When mutant peptide binds to the MHC-I peptide-binding groove, fully assembled peptide-MHC-I complexes (pMHC-I) are presented at the surface of the plasma membrane. The recognition of a neoantigen by a cytotoxic T cell can activate the T-cell response.

To date, most *in silico* approaches for neoantigen prediction have been focused on the MHC-I-related presentation of peptides [6, 7]. Because only 0.5% of peptides can bind to MHC-I molecules, prediction of the binding affinity is the most selective step in the recognition of endogenous antigens [8]. Although these tools provide reliable measurements, practical application at the genome-level is limited. First, the prediction of immunogenic neoantigens primarily relies on a single arbitrary cut-off (50 or 500 nM) of a predicted MHC-I binding affinity. Second, isoform-specific gene expression levels of putative neoantigens and other immune signature-related genes are mostly not estimated or considered. Third, the actual analysis requires a series of complex computational processes, which should be handled by bioinformatics experts manually.

Here, we describe a new machine-learning-based method to address the noted problems above and its implementation. The program, called Neopepsee, harnesses gene- and protein-level information and a novel feature to achieve maximum accuracy: improved performance was validated in a cross validation and an independent *in vivo* study.

Materials and methods

Collection of potential immunogenicity features

To build a classifier that maximizes the usage of information, we initially collected 13 potential features that have been previously reported or hypothesized for predicting immunogenicity. The 13 potential features were divided into three groups based on their representative biologic meanings: (A) MHC-I binding and presentation, (B) amino-acid characteristics, and (C) complex scores. For (A), seven features have been collected and scored, including 'IC₅₀' and 'percentile rank'-based predictive values for MHC-I binding affinity, NetCTLpan [9] based scores for 'MHC-I binding, protein cleavage, TAP transport efficiency' and their 'combined score'. A score for 'T-cell recognition' of the pMHC-I molecule [10] was also considered. For (B), four features have been collected including 'hydrophobicity' of amino acids at TCR contact residues [11, 12], 'polarity and charged value' of amino acids at position 2, 3, 5, and 6 [12], 'molecular size' [13], and 'entropy of peptides' [14]. In addition, two complex scores for 'differential agretopicity index (DAI)' [15] and 'amino acid pairwise contact potentials (AAPPs)' [16] were further considered. More detailed evidences and rationales for collected features are provided in the [supplementary data](#), available at *Annals of Oncology* online.

In addition to the 13 feature, we defined a new score, 'a sequence similarity to known epitopes', in light of a previous hypothesis between the similarity and immunogenicity [17, 18]. As the fundamental objective of the immune response is to distinguish non-self peptides from self, we assumed that these mutations could be prioritized by developing a proper measure. The 14 features are enlisted in the [supplementary Table S1](#), available at *Annals of Oncology* online.

Construction of control positive and negative dataset

To construct the positive dataset ([supplementary Table S2](#), available at *Annals of Oncology* online), we initially collected 1113 epitopes and their corresponding HLA alleles that were reported to exhibit positive T-cell

response in humans [10]. To find epitopes that can be generated by hypothetical somatic mutations in human genome, we compared the sequences of the 1113 epitopes with those of the 20 198 reviewed Swiss-Prot human proteins [19] to only retain 311 epitopes whose sequences are differed from the best-match by up to 2 amino acids; the best-match protein was further used as a corresponding wild-type.

For negative dataset, we initially collected 22 245 variants from common (minor allele frequency [MAF] ≥ 0.05) non-synonymous single nucleotide polymorphisms (SNPs) from dbSNP v.141 with the assumption that widespread peptide variants would not lead to an immunogenic response. The HLA allele for negative dataset was randomly selected from HLA alleles of the positive set. To maintain the naturally occurring balance between positive and negative neopeptides (ratio 1–48) [20], we randomly selected 14 633 out of the 22 245 mutant peptides to finalize the negative dataset.

Feature selection procedure

To achieve better classification accuracy and to reduce the risk of overfitting, we conducted feature selection from the 14 scores. Cleavage and TAP of the 14 scores were initially excluded due to the lack of consistency reported by a previous study [9]. For 12 remaining features, two major criteria were applied for selection: (1) informativeness of the feature in classification and (2) inter-dependency between features that causes redundancy.

To measure informativeness, three different analyses were used: correlation-based feature selection [21], information gain-based feature selection [22], and classification power achieved by a single feature. The single feature based classification power enables the prediction of the discriminative power of a feature in a trained machine learning classifier, which is represented by area under a curve (AUC) and area under a precision-recall curve (AUPRC). The final informativeness is measured by a merged score ([supplementary Figure S1A](#) and [Table S3](#), available at *Annals of Oncology* online). Pearson and Spearman correlation were used to remove overlapping effects between inner individual features. If correlation between two features was higher than 0.8 in both methods, one was excluded. Finally, nine immunogenicity features: IC₅₀, rank, combined score, immunogenicity score, hydrophobicity, polarity&charged score, DAI, AAPPs, and similarity were utilized to construct the machine-learning classifiers (see details in the [supplementary data and Figures S2–S6](#), available at *Annals of Oncology* online).

Machine-learning-based classification for immunogenicity prediction

On the basis of the selected immunogenicity features, the machine-learning-based classifiers were constructed for four learning models: Gaussian naïve Bayes (GNB), locally weighted naïve Bayes (LNB), random forest (RF), and support vector machine (SVM).

To evaluate classifiers, we used 500 iterations of 10-fold cross-validation on the constructed dataset and measured the performance. In the training step, calculated immunogenicity features of the dataset were fed into four classifiers. Then we validated the performance of four trained classifiers and conventional methods on the independent dataset.

Results

Overall workflow of Neopepsee

The overall workflow is shown in Figure 1B. RNA-seq data and a list of somatic mutations are required for analysis. The HLA allele can be supplied as an optional input; otherwise, it can be inferred computationally from the RNA-seq data. For each non-synonymous somatic mutation, affected peptides are calculated

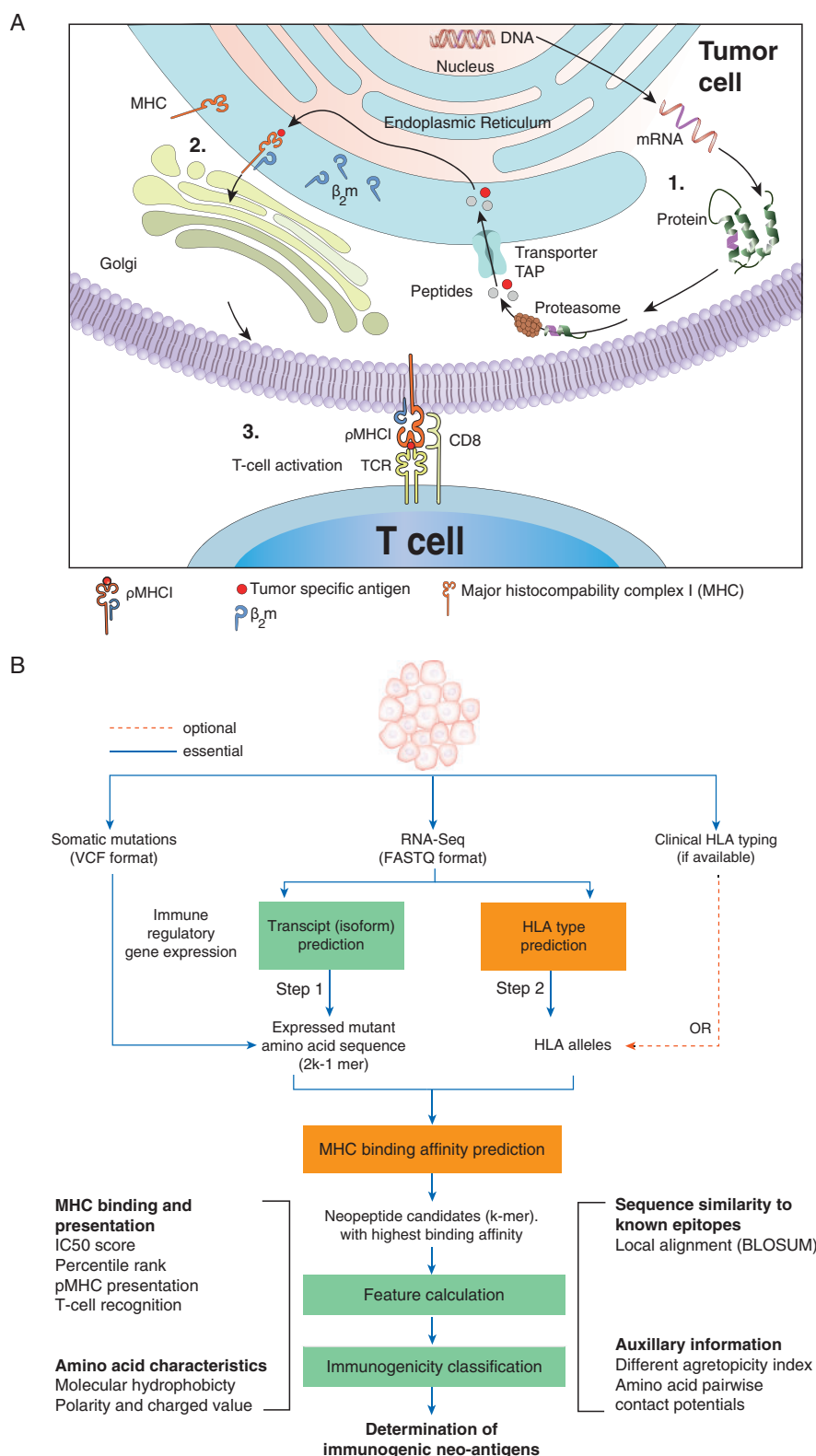


Figure 1. MHC-I antigen presentation pathway and corresponding Neopepsee procedure. (A) The mutant peptide (mutation sequence represented in red) from a tumor-specific antigen can be present in the plasma membrane of the major histocompatibility complex (MHC)-I molecule as the result of a series of reactions. First, tumor-specific antigens are degraded by the proteasome (Process 1). Then, the resulting peptides are transported via transporters associated with antigen presentation (TAP) into the endoplasmic reticulum (ER) lumen. The neopeptide binds to the binding-groove of MHC-I molecules (Process 2). Peptide-MHC class I complexes (pMHC-I) are then transported via the endoplasmic reticulum (ER)-Golgi pathway to the plasma membrane for antigen presentation to activate CD8+ T cells. MHC-I, major histocompatibility complex I; $\beta 2m$, $\beta 2$ -microglobulin; TCR, T-cell receptor. (B) Overall Neopepsee workflow. Somatic mutations and gene expressions of a given tumor tissue are assessed for neoantigen prediction. In total, 9 immunogenicity features are calculated for each mutant neopeptide and fed into a classifier. Green boxes denote unique modules in Neopepsee.

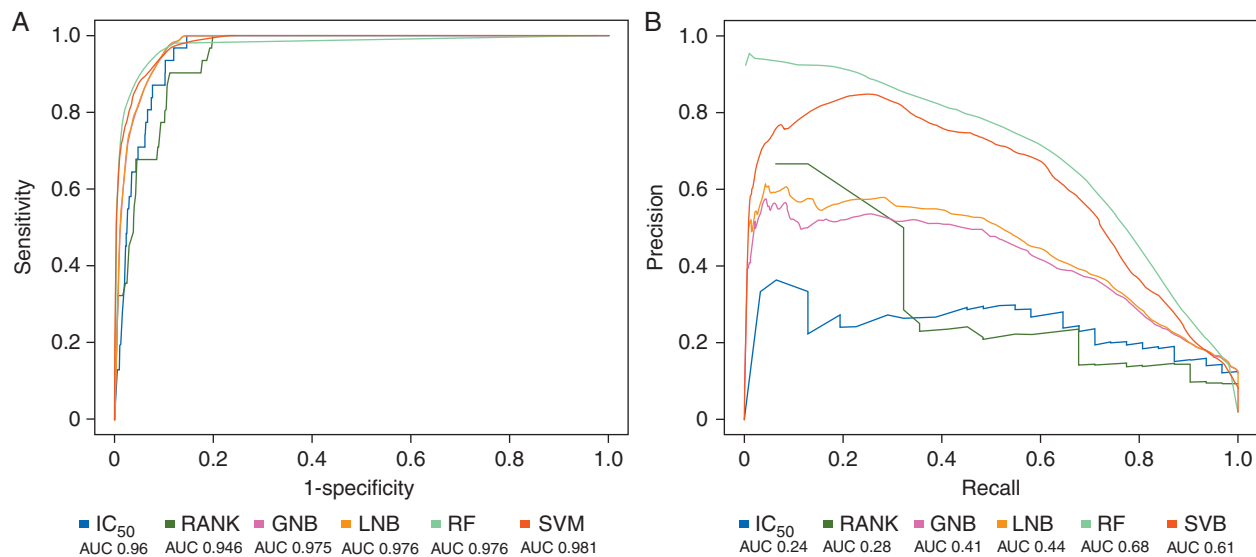


Figure 2. Performance evaluation for Neopepsee. (A) ROC curves of four predictive models (Gaussian naïve Bayes, locally weighted naïve Bayes, random forest, and support vector machine) and two conventional methods (IC₅₀ and percentile rank). All four predictive models outperformed the conventional methods. (B) pROC curves of the same six models. All four predictive models show better performance than the conventional methods. AUPRC is maximized in the RF model; however, an acute drop in the high recall area is observed; LNB shows the best performance when the recall was 1. GNB, Gaussian naïve Bayes; LNB, locally weighted naïve Bayes; RF, random forest; SVM, support vector machine; RANK, percentile rank.

and prepared for further analysis (see details in the [supplementary data](#), available at *Annals of Oncology* online). Neopepsee classifies the potential neoantigens into three categories (high, medium, and low) with respect to the predicted immunogenicity, and reports the results along with 87 immunogenicity-related values, including conventional predictions ([supplementary Table S4](#), available at *Annals of Oncology* online), expression levels of neopeptides and immune regulatory genes (e.g. PD1 and PD-L1).

Evaluation of neoantigen prediction in Neopepsee

Two types of descriptive statistics have been used to analyze the prediction accuracy of machine-learning classifiers (Figure 2). First, the overall AUC of the ROC curve was improved in Neopepsee (GNB: 0.975; LNB: 0.976; RF: 0.976; SVM: 0.981) compared with conventional measures (IC₅₀: 0.96; percentile rank: 0.946) in Figure 2A. Second, AUPRC (Figure 2B) was improved approximately two- to threefold (GNB: 0.41; LNB: 0.44; RF: 0.68; and SVM: 0.61) compared with the IC₅₀ (0.24) and percentile rank (0.28). Overall, the results demonstrate that machine learning-based classifiers have higher classification power than conventional single-value-based approaches.

The overall accuracy at the specific threshold was measured. Conventional thresholds for immunogenicity prediction used 500 nM for IC₅₀ and 2.49 for percentile rank, respectively [2, 18, 23]. We divided the candidate neoantigens into three classes regarding the output membership probability of the trained classifiers: high (the most precise), medium (the most sensitive) and low. The lowest membership probability, which qualifies sensitivity level of 0.95 to exclude outliers, is set as threshold for medium class. The threshold for high class is the highest membership probability at which specificity is above 0.95. All other peptides were classified as the low class. As expected, high class had greatly

increased precision and specificity with a slight loss of sensitivity in the test step, whereas the medium class was optimized for increased sensitivity.

In terms of AUC and AUPRC, the RF and SVM classifiers showed the best performance. However, the low thresholds at the membership probability in the medium class implied the perceptual tendency to call a smaller number of answers to prioritize specificity in both models. The tendency led to a drastic drop of precision at obtaining higher recall (Figure 2B). Therefore, LNB was finally selected as a classification model based on performance robustness.

Tests on independent experimental data

Neopepsee was tested on independent datasets from recent studies that reported experimentally confirmed immunogenic and non-immunogenic peptides in melanoma [23] and chronic lymphocytic leukemia [24] patients (Figure 3). Both studies validated the induced T-cell responses experimentally using a MHC dextramer assay to evaluate vaccine-induced T-cell responses [23] or an IFN- γ ELISPOT assay to measure spontaneous T-cell responses [24]. From both studies, 1093 peptides were obtained, 65 of which were validated (12 immunogenic and 53 non-immunogenic). The number of total calls and true/false positives was assessed in the medium and high classes of Neopepsee and further compared with the conventional criteria (Figure 3A and [supplementary Table S5](#), available at *Annals of Oncology* online).

Regarding sensitivity, IC₅₀ and the medium class called all 12 answers: note that the perfect sensitivity for IC₅₀ is expected, because the peptides for validation were initially selected by IC₅₀. Classification using percentile rank and the high class missed one and two answers, respectively. However, the high class only

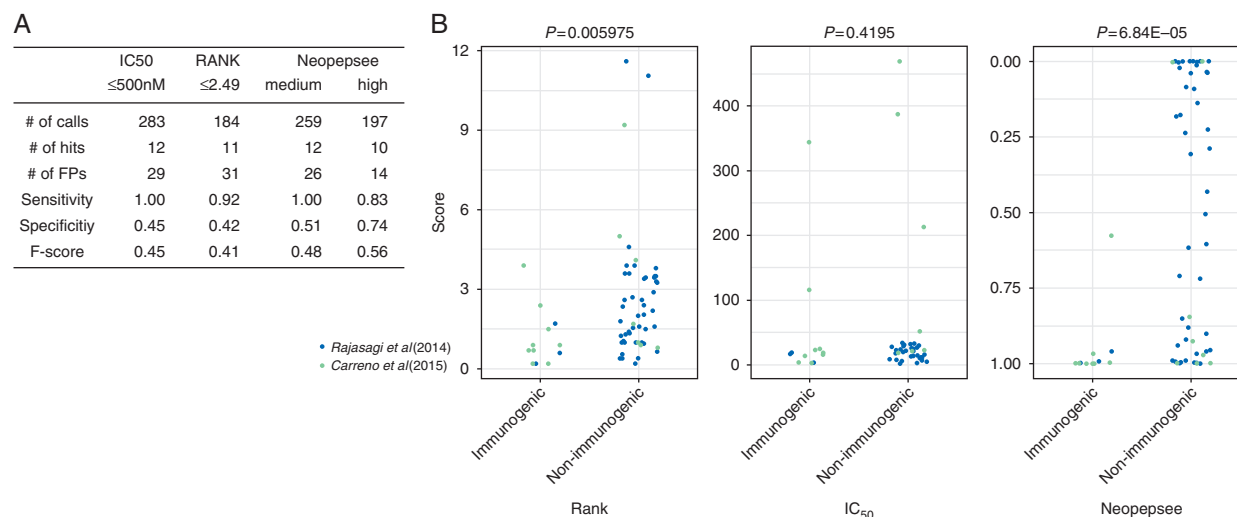


Figure 3. Performance of Neopepsee in independent validation datasets. (A) Comparison of performances between each method. # of calls, the total neoantigen candidates from the each method; # of hits, the number of true positives; # of FPs, the number of false positives. (B) Comparison of discriminative power for three methods. Only Neopepsee shows significance between immunogenic and non-immunogenic scores. The y-axis of Neopepsee was vertically inverted for convenient comparison.

misclassified 14 non-immunogenic peptides out of the 53, increasing in specificity to 0.74 (compared with 0.45 of IC₅₀ and 0.42 of percentile rank). The balanced measure (*f*-score) confirmed the improved classification power (0.41–0.45 in conventional criteria versus 0.48–0.56 in Neopepsee).

For more accurate comparison of discriminative power, the distributions of scores were plotted for immunogenic and non-immunogenic peptide groups (Figure 3B). The *P*-value of discriminative power was calculated by Wilcoxon rank sum test. The results showed a tighter clustering of immunogenic peptides, with a better separation from non-immunogenic peptides (*P*-value, 6.84e–05). While the majority of non-immunogenic peptides are assigned with low probabilistic scores (<0.5), separation of false negatives with very high scores (~1.0) would be the key to achieving better separation, which we expect with continued accumulation of validated data for training.

Application to The Cancer Genome Atlas stomach adenocarcinoma (TCGA-STAD) dataset

We applied Neopepsee to a large cancer genome cohort. A total of 224 samples were analyzed, and 3760 putative neoantigens were identified. The median of somatic single-nucleotide variants (SNV) was 49 and the median of putative neoantigens was 7 (Figure 4A). The number of somatic SNVs largely differed according to microsatellite instability (MSI) status (median of 452 in MSI-high, 56 in MSI-low, and 40 in microsatellite stable [MSS] tumors; Figure 4B). The number of neoantigens in MSI-high tumors (median, 68) was higher than that of MSI-low (median, 10) or MSS tumors (median, 5). Interestingly, tumors with neoantigens exhibited better prognosis compared with tumors without neoantigens (29.1 versus 14.1 months without neoantigens; log-rank *P*=0.024; Figure 4C). However, MSI status was not significantly associated with overall survival (29.4 versus 26.7 months with MSS; log-rank *P*=0.616; Figure 4D). Recently, Rooney et al. calculated immune cytolytic activity scores as

indicators of CD8+ T-cell activation and showed neoantigens to be likely to induce cytolytic activity [25]. In the cox-regression analysis, the absence of neoantigens and advanced tumor stage (III and IV) were identified as independent prognostic factors for overall survival (Figure 4E). The positivity of neoantigen was associated with improved overall survival in both univariate (hazard ratio, 3.1; *P*=0.022) and multivariate analysis (hazard ratio, 2.2; *P*=0.040). Our data suggest that increased neoantigen loads by microsatellite instability, not MSI status itself, may be a favorable prognostic factor. Of 3760 neoantigens, only 16 (0.42%) were found in more than one tumor sample (supplementary Table S6, available at *Annals of Oncology* online). We compared identified neoantigens with known immune epitopes. In addition to known *Helicobacter pylori* epitopes (*n*=7, Supplementary Figure S7A, available at *Annals of Oncology* online), we identified a total of 1867 known immune epitopes (supplementary Table S7, available at *Annals of Oncology* online). Most immune epitopes are derived from *Mycobacterium tuberculosis* (24%), *Trypanosoma cruzi* (19%), *Vaccinia virus* (6%), Human herpes virus (4%), and, Hepatitis C virus (2%), many of these bacteria/viruses can induce chronic inflammation, one of critical mediators of tumor development including gastric cancer (supplementary Figure S7B, available at *Annals of Oncology* online).

Discussion

Owing to the great diversity, multi-step biochemical processes, and the stochastic nature of T-cell immune response, accurate prediction of neoantigen has been one of the most challenging problems in the immunoinformatics field. As we assumed, a single value associated with a part of the whole process can hardly provide sufficient information to resolve the complexity. In this context, systematic integration of multiple features into a unified workflow is urgently needed to increase the accuracy and to

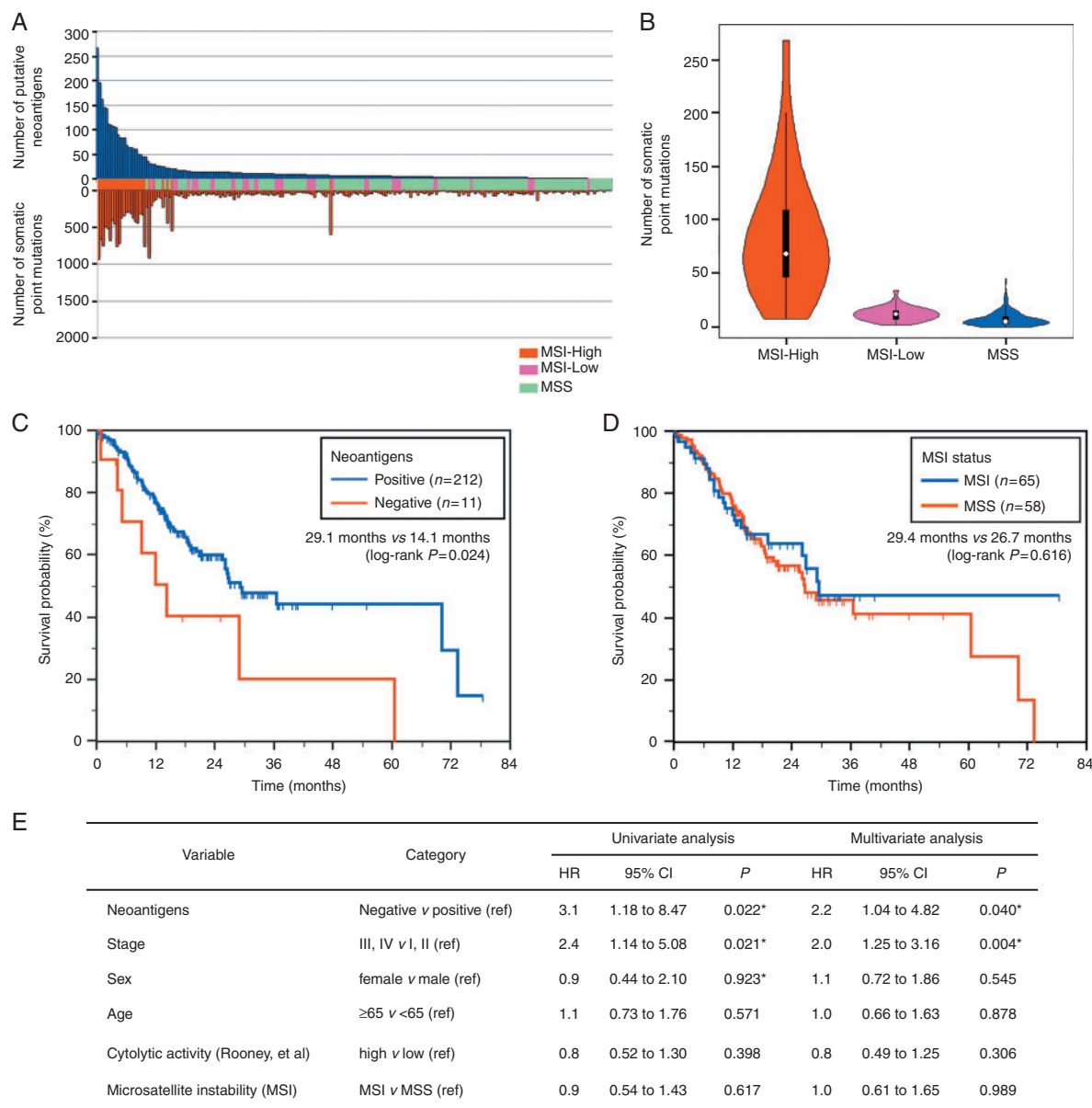


Figure 4. Application of Neopepsee to stomach adenocarcinoma (TCGA-STAD) dataset. (A) The number of putative neoantigens and somatic point mutations in TCGA-STAD tumors (N=224). (B) The number of somatic point mutations according to microsatellite instability (MSI) status in TCGA-STAD tumors (N=224). (C) Kaplan–Meier estimates of overall survival according to the positivity of neoantigens. (D) Kaplan–Meier estimates of overall survival according to the MSI status. (E) Univariate and multivariate Cox regression survival analyses.

provide a sustainable framework that exploits the growing information. Notably, a recent study to predict MHC-I-binding peptides based on mass spectrometry shed light on identifying binding motifs and antigen processing rules [26]. We expect that such advances will reinforce Neopepsee through a continuous re-training of the machine learning classifier with updated datasets and feature selection.

One of the ultimate goals of neoantigen prediction software is to classify patients who will benefit from immunotherapy, or to design a personalized cancer vaccine. Recent studies showed that higher mutational burden was correlated with better antitumor activity of CTLA4 or PD-1 blockade [3, 4, 18]. Although the number of neoepitopes generated by somatic mutation seemed to be important for predicting antitumor activity of immunotherapy, the

criteria to identify neoepitopes and the correlation were inconsistent across studies [3, 18]. Future studies identifying optimized selection criteria for neoepitopes and further correlative analysis are warranted. Since most neoantigenic peptides are not identical between tumors [4], cancer vaccines targeting neoepitopes may not be a ‘universal’ solution that provides broad coverage to cancer patients. Nevertheless, Neopepsee will enable the efficient analysis of a personal somatic mutation profile and identification of potential neoepitopes for personalized vaccination.

In summary, Neopepsee can be applied not only to identify putative neoantigens, but also to compare neoantigens with known immune epitopes. The analysis results can be used for subsequent prognostic/predictive biomarker discovery or to design antigens for cancer vaccines.

Availability

The program is available at <http://sourceforge.net/projects/neo pepsee/>.

Acknowledgements

The authors thank Dong-Su Jang (Medical Illustrator, Department of Research Affairs, Yonsei University College of Medicine, Seoul, South Korea) for his help in creating the medical illustrations.

Funding

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2015R1C1A1A01053638); HSK and MGL were supported by the National Research Foundation, the Ministry of Science, ICT & Future Planning (2013R1A3A2042197); and the Korea Health Technology R&D Projects through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare (HI14C1324), Republic of Korea. SWK was additionally funded by a faculty research grant from the Yonsei University College of Medicine (6-2016-0081).

Disclosure

The authors have declared no conflicts of interest.

References

1. York IA, Rock KL. Antigen processing and presentation by the class I major histocompatibility complex. *Annu Rev Immunol* 1996; 14(1): 369–396.
2. Brown SD, Warren RL, Gibb EA et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res* 2014; 24(5): 743–750.
3. Van Allen EM, Miao D, Schilling B et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015; 350(6257): 207–211.
4. Rizvi NA, Hellmann MD, Snyder A et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015; 348(6230): 124–128.
5. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 2011; 11(12): 823–836.
6. Vita R, Overton JA, Greenbaum JA et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015; 43(Database issue): D405–D412.
7. Hoof I, Peters B, Sidney J et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009; 61(1): 1–13.
8. Trost B, Bickis M, Kusalik A. Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. *Immunome Res* 2007; 3(1): 5.
9. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 2010; 62(6): 357–368.
10. Calis JJ, Maybeno M, Greenbaum JA et al. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* 2013; 9(10): e1003266.
11. Chowell D, Krishna S, Becker PD et al. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc Natl Acad Sci USA* 2015; 112(14): E1754–E1762.
12. Patronov A, Doytchinova I. T-cell epitope vaccine design by immunoinformatics. *Open Biol* 2013; 3(1): 120139.
13. Dintzis HM, Dintzis RZ, Vogelstein B. Molecular determinants of immunogenicity: the immunon model of immune response. *Proc Natl Acad Sci USA* 1976; 73(10): 3671–3675.
14. Liu MK, Hawkins N, Ritchie AJ et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J Clin Invest* 2013; 123(1): 380–393.
15. Duan F, Duitama J, Al Seesi S et al. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J Exp Med* 2014; 211(11): 2231–2248.
16. Saethang T, Hirose O, Kimkong I et al. PAAQD: predicting immunogenicity of MHC class I binding peptides using amino acid pairwise contact potentials and quantum topological molecular similarity descriptors. *J Immunol Methods* 2013; 387(1–2): 293–302.
17. Hoof I, Perez CL, Buggert M et al. Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. *J Immunol* 2010; 184(9): 5383–5391.
18. Snyder A, Makarov V, Merghoub T et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 2014; 371(23): 2189–2199.
19. Apweiler R, Bairoch A, Wu CH et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004; 32(Database issue): D115–D119.
20. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* 2015; 348(6230): 69–74.
21. Hall MA. Correlation-based feature selection of discrete and numeric class machine learning (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science; 2000.
22. Azhagusundari B, Thanamani AS. Feature selection based on information gain. *Int J Innov Technol Exploring Eng (IJITEE)* 2013; 2: 18–21.
23. Carreno BM, Magrini V, Becker-Hapak M et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* 2015; 348(6236): 803–808.
24. Rajasagi M, Shukla SA, Fritsch EF et al. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* 2014; 124(3): 453–462.
25. Rooney MS, Shukla SA, Wu CJ et al. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 2015; 160(1–2): 48–61.
26. Abelin JG, Keskin DB, Sarkizova S et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 2017; 46(2): 315–326.