High-throughput prediction of MHC Class I and Class II neoantigens with MHCnuggets

Shao XM[1,2*], Bhattacharya R[1,3 *],Huang J[1,3 *], Sivakumar IKA[1,3,4*], Tokheim C[1,2], Zheng L[1,5], Kaminow B[1,6], Omdahl A[1,2], Bonsack M[7,8,9], Riemer AB[7,8], Velculescu VE[1,5, 10], Anagnostou V[10], Pagel KA[1,2], Karchin R[1,2,10⇤]

1 Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA

2 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

3 Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

4 Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA

5 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

6 Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA

7 Immunotherapy and Immunoprevention, German Cancer Research Center (DKFZ), Heidelberg, Germany

8 Molecular Vaccine Design, German Center for Infection Research (DZIF), partner site Heidelberg, Heidelberg, Germany

9 Faculty of Biosciences, Heidelberg University, Heidelberg, Germany

10 The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA

* These authors contributed equally
⇤Corresponding author: Rachel Karchin, Ph.D
217A Hackerman Hall
3400 N. Charles St. Baltimore, MD USA 21204 ph: +1 410 516 5578
fax: +1 410 516 5294 karchin@jhu.edu

Running title: High-throughput prediction of neoantigens with MHCnuggets

Potential Conflicts of Interest: V.A receives research funding from Bristol-Myers Squibb. V.E.V. is a founder of Personal Genome Diagnostics, a member of its Scientific Advisory Board and Board of Directors, and owns Personal Genome Diagnostics stock, which are subject to certain restrictions under university policy. V.E.V. is an advisor to Takeda Pharmaceuticals. Within the last five years, V.E.V. has been an advisor to Daiichi Sankyo, Janssen Diagnostics, and Ignyta. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies.

Abstract

Computational prediction of binding between neoantigen peptides and major histocompatibility complex (MHC) proteins is an emerging biomarker for predicting patient response to cancer immunotherapy. Current neoantigen predictors focus on *in silico* estimation of MHC binding affinity and are limited by low positive predictive value for actual peptide presentation, inadequate support for rare MHC alleles and poor scalability to high-throughput data sets. To address these limitations, we developed MHCnuggets, a deep neural network method to predict peptide-MHC binding. MHCnuggets is the only method to handle binding prediction for common or rare alleles of MHC Class I or II, with a single neural network architecture. Using a long short-term memory network (LSTM), MHCnuggets accepts peptides of variable length and is capable of faster performance than other methods. When compared to methods that integrate binding affinity and HLAp data from mass spectrometry, MHCnuggets yields a fourfold increase in positive predictive value on independent MHC-bound peptide (HLAp) data. We applied MHCnuggets to 26 cancer types in TCGA, processing 52.6 million allele-peptide comparisons in under 2.3 hours, yielding 103,587 candidate immunogenic missense mutations (IMMs). IMM hotspots occurred in 36 genes, including 22 driver genes. Predicted IMM load was significantly associated with increased immune cell infiltration (p<2e-16) including CD8+ T cells. Notably, only 0.15% of predicted immunogenic missense mutations were observed in >2 patients, with 65% of these derived from driver mutations. Our results provide a new method for neoantigen prediction with high performance characteristics and demonstrate its utility in large data sets across human cancers.

Synopsis

We developed a new *in silico* predictor of Major Histocompatibility Complex (MHC) ligand binding and demonstrated its utility to assess potential neoantigens and immunogenic missense mutations (IMMs) in 6613 TCGA patients.

Introduction:

The presentation of peptides bound to major histocompatibility complex (MHC) proteins on the surface of antigen-presenting cells and the subsequent recognition by T-cell receptors is fundamental to the mammalian adaptive immune system. Recent advances in cancer immunotherapy have brought mutation-associated neoantigens to the epicenter of tumor response and have highlighted the need for improved understanding of which peptides will bind to MHC proteins and generate an anti-tumor immune response (1-4). Neoantigens derived from somatic mutations have been shown to be targets of immunoediting and drive therapeutic responses in cancer patients treated with immunotherapy (5,6). Due to the fact that experimental characterization of neoantigens is both costly and time-consuming, many computational methods have been developed to predict peptide-MHC binding and the subsequent immune response (7,8). To date, supervised neural network machine learning approaches are the best-performing (9-11) and the most widely used *in silico* methods for this purpose. Despite these advances, computational approaches using modern neural network architectures have been unable to significantly improve predictive performance in the past several years, due in part to lack of sufficiently large sets of experimentally characterized peptide binding affinities for most MHC alleles.,

While neoantigen prediction for common MHC Class I alleles is well-studied (12), predictive accuracy on rare and less well-characterized MHC alleles remains poor (13-15) and there is a general scarcity of Class II predictors (16). Current estimates suggest that Class II antigen lengths primarily range from 13-25 amino acids (17), and this diversity has been a major obstacle to developing *in silico* neoantigen predictors (16,18). As most neural network architectures are designed for fixed-length inputs, methods such as NetMHC (19-22)and MHCflurry (23) require pre-processing of peptide sequences or extensive training of separate classifiers for each peptide length.

Clinical application of MHC-peptide binding predictors, to identify biomarkers for cancer immunotherapy, requires predictors that are scalable to large patient cohorts with low false positive rates (24). A cancer may contain hundreds of candidate somatically altered peptides, but few will actually bind to MHC proteins and elicit an immune response (25). For many years, most neoantigen predictors were trained primarily on quantitative peptide-HLA binding affinity data from *in vitro* experiments (26). More recently, advances in immunopeptidomics technologies have enabled identification of thousands of naturally presented MHC bound peptides (HLAp) from cancer patient samples and cell lines (27) (28) (24). The potential to improve neoantigen predictors by integrating binding affinity and HLAp data (24) has motivated new hybrid approaches (19,23). Despite these advances, most methods predict large numbers of peptides as candidate neoantigens, of which only a few are actually immunogenic in patients

(16,24,29).

In this work, we present a long short-term memory (LSTM) neural network method, MHCnuggets, as the first neoantigen predictor designed for MHC Class I and Class II alleles in a single framework. The method leverages transfer learning and allele clustering to accommodate both common, well-characterized MHC alleles and less studied rare alleles. As computational neoantigen predictors generate a large ranked list of candidate peptides, we reasoned that maximizing the number of highly-ranked true positives would be preferred in many applications (23). We demonstrate competitive predictive performance of MHCnuggets to widely-used methods on binding affinity datasets. In comparison to hybrid methods that have integrated binding affinity and HLAp data, we show decreased false positives and increased positive predictive value in a held-out cell line data set of ligands identified by mass spectrometry (11,30). To demonstrate the clinical utility and scalability of MHCnuggets to large patient cohorts, we investigated candidate immunogenic mutations from 26 tumor types in The Cancer Genome Atlas (TCGA). MHCnuggets yielded 103,587 candidate immunogenic missense mutations (out of 1,124,266) in less than 2.3 hours. These mutations were correlated with increased lymphocyte infiltration, however only 0.15% were observed in multiple patients.

Methods:

Implementation

MHCnuggets uses a long short-term memory (LSTM) neural network architecture (31) (Figure 1A). LSTM architectures excel at handling variable length sequence inputs and can learn long-term dependencies between non-contiguous elements, enabling an input encoding that does not require peptide shortening or splitting (Figure 1B).   The networks were trained with a transfer-learning protocol (32), which allows networks performing predictions for less well-characterized alleles to leverage information from extensively studied alleles (Figure 1C). Transfer learning was also used to train networks using both binding affinity and HLAp datasets. In addition, MHCnuggets' architectures can be trained using either continuous binding affinity measurements from *in vitro* experiments (half maximal affinity or IC50) and/or immunopeptidomic (HLAp) binary labels. The former utilizes a mean-squared error (MSE) loss while the latter utilizes binary cross-entropy (BCE) loss for training.

For each MHC allele, we trained a neural network model consisting of an LSTM layer of 64 hidden units, a fully connected layer of 64 hidden units and a final output layer of a single sigmoid unit. For the 16 alleles where allele-specific HLAp training data was available (33), we trained networks on both binding affinity and HLAp data (MHCnuggets). Next, we trained networks only with binding affinity measurements (MHCnuggets noMS) for all MHC Class I alleles. Due to the lack of allelic-specific HLAp training data for Class II, all MHC Class II networks were trained only on binding affinity measurements. In total, we trained 148 Class I and 136 Class II allele-specific networks. Common alleles with many characterized binding peptides comprise a small fraction of all known MHC alleles (34). To handle binding predictions for rare alleles, MHCnuggets selects a network by searching for the closest allele, based on previously published supertype clustering approaches (35,36). Briefly, HLA-A and HLA-B alleles were clustered by MHC binding pocket amino acid residue composition, and HLA-C and all MHC II alleles were hierarchically clustered based upon experimental mass spectrometry and binding assay results. For alleles with no supertype classification, the closest allele was from the same HLA gene, and allele group if available, with preference for alleles with the largest number of characterized binding peptides. All networks were implemented with the Keras Python package (TensorFlow back-end) (37,38). The open source software is available at https://github.com/KarchinLab/mhcnuggets-2.3, installable via pip, and has been integrated into the PepVacSeq (39) and Neoepiscope (40) pipelines.

Benchmarks

To accurately assess the performance of MHCnuggets on a variety of MHC-peptide binding prediction tasks, we utilized six distinct benchmark sets (Table S1). The benchmarks were designed to evaluate binding prediction for MHC Class I alleles, MHC Class II alleles, well-characterized alleles with a trained model (allele-specific prediction) and rare alleles with limited or no experimental peptide binding data (pan-allele prediction) (Figure 2). To compare to the widely-used HLA ligand prediction tools from the NetMHC group (NetMHC3.0, NetMHC 4.0, NetMHCpan2.0, NetMHCpan 4.0) (21,22), which incorporate IEDB data and can be trained only by their developers, as well as the open source MHCflurry tools, we decided to employ multiple benchmarking strategies. The four strategies include: 1) an independent benchmark test set of peptides not included as training data for any of the methods; 2) a previously published paired training/testing benchmark; 3) a five-fold cross-validation benchmark; 4) leave-one-molecule-out (LOMO) benchmark.

We evaluated six MHC Class I predictors on independent binding affinity and HLAp datasets (11,12,30). First, we compared MHCnuggets to several Class I predictors that incorporate both binding affinity and HLAp data: MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. Each method was benchmarked using an independent set of MHC-bound peptides identified by mass spectrometry across seven cell-lines for six MHC I alleles (Bassani-

Sternberg 2017, Trolle 2016). For testing, HLAp hits were combined with random decoy peptides sampled from the human proteome (33) in a 1:999 hit-decoy ratio, totaling 26,317,000 peptides. Next, four MHC Class I predictors trained only on binding affinity data (MHCnuggets (noMS) and MHCflurry (noMS), NetMHC 3.0 and NetMHCpan 2.0) were evaluated with the Kim et al dataset (9), in which each predictor was trained with the BD2009 data and tested on BLIND data. It was possible to compare NetMHC3.0 and NetMHCpan2.0 performance on Kim et al., because they have previously published predicted IC50s for all peptide-MHC pairs in BLIND. This allowed us to calculate their PPV_n, area under the ROC curve (auROC), Kendall's *tau*, and Pearson's *r* correlations.

Next, we compared MHCnuggets to the MHC Class II ligand prediction methods from the NetMHC group (41). Such comparison was only possible through their self-reported summary performance statistics. We used the Jensen *et al*. five-fold cross-validation benchmark to assess allele-specific MHC Class II prediction of MHCnuggets and NetMHCII 2.3, for 27 alleles. NetMHCII 2.3 reported the average auROC for five-fold cross-validation, and we report MHCnugget's positive predictive value for each of the 27 alleles as well as the average auROC, Pearson's *r* and Kendall-Tau correlations.

The leave-one-molecule-out (LOMO) benchmarks are a type of cross-validation designed to estimate the performance of peptide binding prediction with respect to rare, poorly characterized MHC alleles, which lack binding affinity training data. Given training data for *n* MHC alleles, the data for a single allele is held out and networks are trained for the remaining *n-1* alleles. Then for each peptide, predictions are generated by the remaining networks. We designed a LOMO benchmark to evaluate MHC Class I rare allele prediction, by selecting 20 alleles with 30 to 100 characterized peptides in IEDB. For Class II rare allele prediction, we used the Jensen et al. LOMO benchmark. We were unable to assess rare allele prediction for NetMHC Class I methods, as no published results were available. For the NetMHC Class II methods, we compared MHCnuggets to their self-reported auROCs.

Runtime analysis

To assess the speed and scalability of the tested methods, we selected one million peptides sampled from the Abelin et al. dataset (33) for Class I alleles, and one million peptides sampled from the IEDB (curated dataset 2018 (42)) for Class II alleles. Sampling was done with replacement. For each method listed in Figure 1A, networks for three Class I MHC alleles (HLA-A*02:01, HLA-A*02:07, HLA-A*01:01) and three Class II MHC alleles (HLA-DRB1*01:01, HLA-DRB1*11:01, HLA-DRB1*04:01) were used to predict binding over a range of input sample sizes (10^2, 10^3, 10^4, 10^5, 10^6). All methods were run on a single GPU compute node (one NVIDIA TESLA K80 GPU plus six 2.50GHz Intel Xeon E5-2680v3 CPUs, 20GB memory).

TCGA analysis pipeline

To assess candidate immunogenic somatic mutations in patients from the TCGA cohort, we developed and implemented a basic pipeline based on whole-exome and RNA sequencing data. Our analysis builds upon work from the TCGA PanCancer Analysis teams for drivers (43), mutation calling (44) and cancer immune landscapes (45). We obtained somatic mutation calls for all cancer types from Multi-Center Mutation Calling in Multiple Cancers (MC3) (v0.2.8) (7775 patients). Tumor-specific RNA expression values from Broad TCGA Firehose were standardized across tumor types using the RSEM Z-score (46). MHC allele calls were obtained from the TCGA cancer immune landscape publication, in which up to six MHC Class I alleles (HLA-A, HLA-B, and HLA-C) were identified for each patient using OptiType (47). We included patients for which mutation calls, MHC allele calls and RNA expression values were available from TCGA (Supplementary Methods). After these considerations, the analysis included 6613 patients from 26 TCGA tumor types. Six cancer types were not included in our analysis, because 15 or fewer patients met this requirement: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Esophageal carcinoma (ESCA) , Mesothelioma (MESO), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Ovarian serous cystadenocarcinoma (OV).

The somatic missense mutations identified in each patient were filtered to include only those with strong evidence of mutant gene RNA expression in that patient (Z>=1.0). For each mutation that passed this filter, we used the transcript assigned by MC3 to pull flanking amino acid residues from the SwissProt database (48), yielding a 21 amino acid residue sequence fragment centered at the mutated residue. All peptides of length 8,9,10 and 11 that included the mutated residue were extracted from each sequence fragment. Next binding affinity predictions were generated for each mutated peptide and its paired germline peptide for up to six MHC Class I alleles, depending on the patient's HLA genotypes. In total, each somatic mutation was represented by 38 mutated peptides and 38 matched germline peptides for up to 6 possible MHC pairings.

We applied a permissive filter to select candidate immunogenic peptides, requiring that at least one MHC allele was predicted to have binding affinity of IC50<500nM with respect to the mutated peptide and that the reference peptide had IC50 at least two-fold larger. The ratio of peptide binding affinity of a germline peptide to its matched mutated peptide, known as differential agretopic index (DAI) has been previously shown to be a strong signal of point mutation immunogenicity (40,49,50). Somatic missense mutations that generated neoantigens meeting these criteria were considered candidate immunogenic missense mutations (IMMs). If multiple neoantigens and/or MHC alleles were predicted by MHCnuggets, the IMM was counted only once for the MHC allele with maximum DAI. Finally, for each patient we counted the number of candidate IMMs found in their exome and stratified by tumor type. We then identified candidate IMMs that were harbored by more than one patient.

We sought to ascertain whether candidate IMMs occurred preferentially in particular gene or protein regions. To assess the largest number of protein sequences, we used clustering in primary amino acid residue sequence to identify statistically significant IMM hotspots (HotMaps 1D algorithm v1.2.2) (51) q<0.1, Benjamini-Hochberg method (52)). In this analysis, mutations were stratified by cancer type, and we considered enrichment within linear regions of length 50 amino acid residues.

We considered that mutation immunogenicity might be associated with potential driver status of a mutation. Driver status was inferred by CHASMplus (53), a random forest classifier that predicts driver missense mutations. It utilizes a multi-faceted feature set, incorporating mutational hotspot detection, annotations about molecular function and adjusts for gene-level covariates. It has been previously shown to be effective at identifying both common and rare driver mutations. For each mutation, its immunogenicity was represented as a binary response variable and driver status was used as a covariate. Mutations with CHASMplus q-value < 0.1 were considered drivers (53). We modeled the relationship with univariate logistic regression (R glm package with binomial link logit function).

To assess whether the total number of candidate IMMs per patient was associated with changes in tumor immune infiltrates, we performed Poisson regression (R glm package with Poisson link log function). All estimates of immune infiltrates were obtained from Thorsson et al. (45,54). We fit two univariate models in which the response variable was the IMM count and the covariate was either total leukocyte fraction or fraction of CD8+ T-cells. Our results supported an association between increased IMM load and significant increases in both total leukocyte fraction and CD8 T-cells.

Results:

High-throughput MHCnuggets breaks the MHC ligand prediction plateau

The MHCnuggets LSTM neural network architecture accepts peptides of variable lengths as inputs so that ligand binding prediction can be performed for both MHC Class I and Class II alleles. To enable prediction for peptides that bind to rare MHC alleles with limited experimental data, in addition to several hundred allele-specific networks for common alleles, we designed a method to predict binding to a closely-related common allele. When available, we utilize a transfer learning protocol to integrate binding affinity and HLAp results in a single network model, to better represent the natural diversity of MHC-binding peptides.

To assess the baseline performance assessment for MHCnuggets' allele-specific networks on binding affinity data, we compared our approach to the most widely used MHC Class I ligand prediction methods, using two validation sets of binding affinity measurements (Kim et al. (9) Bonsack et al. (12)). We trained and tested MHCnuggets (noMS) and MHCflurry (noMS) using the Kim *et al* dataset, and evaluated the predictions provided by NetMHC 3.0 and NetMHCpan 2.0. We observed that MHCnuggets' performance (PPV_n = 0.829, auROC=0.924) was comparable to these methods (Figure 3a) (PPV_n of all methods=0.825 +/- 0.005, auROC of all methods = 0.928 +/- 0.0031). MHCnuggets was also comparable (PPV_n = 0.633, auROC=0.794) to these methods when tested on the Bonsack *et al*. dataset (PPV_n of all methods = 0.625 +/- 0.008, auROC of all methods = 0.77 +/- 0.02) (Figure 3A) (+/- refers to standard deviation) (Table S3a, S3b, Table S4a, S4b).

Historically, neoantigen prediction methods have focused on Class I and trained on binding affinity data from IEDB (42). More recent work has incorporated both binding affinity and HLAp data into network training (19,23). We compared MHCnuggets to several Class I predictors that also used both binding affinity and HLAp data: MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. We selected the Bassani-Sternberg/Trolle (BST) HLAp dataset (11,30,33) as an independent benchmark, as it was not previously included as training data by any of these methods. For all alleles tested, MHCnuggets achieved an overall PPV_n of 0.46 and auROC of 0.85 (Figure 3B). On average, MHCnuggets' PPV_n was more than three times higher than MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. For all alleles, MHCnuggets predicted substantially fewer binders than other methods, resulting in fewer false positive predictions. We further analyzed PPV_n performance by stratifying according to peptide length. McNuggets' increased PPV_n was most prominent for peptides of length 9, 10, and 11 (Figure 3C). The length distribution of predicted binders was also commensurate with the observed distribution of naturally occurring binders in the HLAp benchmark tests (Trolle 2016 (Table S5a, S5b, S5c, S5d).

For some clinical applications, it may be desirable to minimize the number of false positives among a small number of top-scored peptides. We also compared PPV of the methods listed above on their top 50 and 500 ranked peptides from the BST dataset (six MHC Class I alleles). MHCnuggets exhibited the highest PPV in the top 50 for all alleles except HLA-B*51:01 and the highest PPV in the top 500 for all alleles (Figure 3D, Table S5e).

Prediction of peptide-MHC binding for Class II and rare alleles

We assessed baseline performance of MHCnuggets Class II allele-specific networks on binding affinity data. To enable comparison with the Class II methods from the NetMHC group, we used a five-fold cross validation benchmark derived from IEDB that was included in the publication describing NetMHCII-2.3 and NetMHCIIpan-3.2 (41). First, we computed PPV_n for each of the 27 allele-specific networks separately (Figure 4A) (mean PPV_n=0.739). Next, we computed the overall auROC, Pearson r and Kendall Tau correlations for all 27 Class II alleles. MHCnuggets overall auROC (0.849) was comparable to that of the NetMHCII-2.3 (0.861) and NetMHCIIpan-3.2 (0.861). Comparison to NetMHC Class II methods was limited to overall auROC as published in (41), because their results are not publicly available (Figure 4B) (Table S6a, Table S6b).

We estimated performance for those Class I and Class II MHC alleles for which we were unable to train allele-specific networks, using leave-one-molecule-out (LOMO) cross-validation (41). In this LOMO protocol, MHC-peptide binding is assessed for a well-characterized allele that has been held out from training, to approximate prediction performance for a rare allele (Figure 5A). For the 20 Class I alleles, the mean PPV_n was 0.65 and the mean auROC was 0.671. For the 27 Class II alleles, the mean PPV_n was 0.65 and the mean auROC was 0.792. In comparison, the Class II mean auROC of NetMHCIIpan-3.2 was 0.781 (Figure 5B, Figure 5C). Further performance results of NetMHCpan rare allele predictors for both Class I and Class II were not publicly available for LOMO tests (Table S7, Table S8a, Table S8b).

Fast and scalable computation

When run on a GPU architecture, MHCnuggets was substantially faster and scaled more efficiently than MHC ligand predictors from the NetMHC family and MHCflurry. Given an input of one million peptides randomly selected from Abelin et al., MHCnuggets runtime was 4.5, 3.2, and 18 times faster than MHCflurry 1.2.0, NetMHC 4.0, NetMHCpan 4.0, respectively (Figure 6A). The improvement was even more pronounced for Class II peptides, for which an input of one million peptides to MHCnuggets ran 65.6 times and 126 times faster than NetMHCII2.3 and NetMHCIIpan 3.2, respectively (Figure 6B). As the total number of input peptides was increased from 0 to one million, the runtime per peptide plateaued for other methods but decreased exponentially for MHCnuggets.

Candidate MHC Class I immunogenic missense mutations in TCGA patients

To illustrate how MHCnuggets' improvements in scalability and positive predictive value could provide utility in the analysis of very large patient cohorts, we developed a basic pipeline to predict Class I MHC-ligand binding in patients sequenced by the TCGA consortium (Methods). As incorporated into the pipeline, patient exomes were split into 21 amino acid residue sequence fragments, centered on each somatic missense mutation. For each 8-, 9-, 10- and 11-length peptide window in the sequence fragment, MHCnuggets predicted the MHC binding of the peptide with the somatic mutation and the binding of the peptide translated from the reference transcript (Ensembl reference transcript from Multi-Center Mutation Calling in Multiple Cancers (MC3) (v0.2.8) (44) translated peptide from SwissProt, using UniProt mapping service (48)). Next, we reduced the very large number of candidate peptides with filters that considered expression and differential binding affinity of somatically mutated peptides compared with reference peptides. We identified candidate *immunogenic missense* mutations (IMMs) as those which generated peptides that passed these filters, for at least one patient-specific MHC allele (Table S9a). Finally, we characterized their predicted driver status and positional hotspot propensity.

Total processing time for 52,569,276 allele-peptide comparisons supported by RNAseq expression was under 2.3 hours. First, we sought to ascertain the extent of variability in predicted IMM count among individuals with different cancer types. Next, we identified IMMs that were shared across patients and protein regions that were highly enriched for IMMs across patients, because these might be candidates for neoantigen-based therapeutic applications. Then we considered whether IMMs were more or less likely to be driver mutations. Finally, we assessed the associations between patient IMM load and computationally estimated immune cell infiltrates.

After applying a strict gene expression filter, we identified 103,587 candidate IMMs in 26 TCGA cancer types, with a mean of 14.9 IMMs per patient. We found that the majority of patients harbored fewer than 10 IMMs, and 900 patients had none. Seventy-six percent of patients had between 1 and 10 IMMs, compared to 1.8% of patients with more than 100, and nine patients with more than 1000 (Figure 7A). Cancer types with the highest number of IMMs were uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD), and lung adenocarcinoma (LUAD), previously known for high mutation burden and immunogenicity (45). UCEC and COAD are also known to have a high frequency of microsatellite-instable (MSI) tumors. The lowest number were found in Uveal Melanoma (UVM), Paraganglioma & Pheochromocytoma (PCPG), and Testicular Germ Cell Cancer (TGCT) (Figure 7B, Table S9b).

Across all cancer types, we identified 1,379 IMMs harbored by two or more patients, of which 157 were identified in three or more patients. Of these 157 only 11.5% occurred exclusively in a single cancer type (Figure 7C). The IMMs identified in the largest number of patients were *IDH1* R132H (62), *PIK3CA* E545K (25), *FGFR3* S249C (23), *PIK3CA* E542K (19), *AKT1* E17K (14), *KRAS* G12D (14) and *KRAS* G12V (14), which are well known recurrent oncogenic driver mutations (55,56).

Genes with the largest number of shared IMMs include *P53* (63), *CTNNB1* (18), *PIK3CA* (15), *KRAS* (8), *HRAS* (8), *PTEN* (6) and *FBXW7* (7), *EP300* (5) and *POLE* (5). Among the few IMMs observed in a single cancer type in three or more patients were *SF3B1* K700E (BRCA), *EGFR* L858R (LUAD), *CIC* R215W (LGG), *GNA11* Q209L and *SF3B1* R625H (UVM) and *NFE2L2* R34G (BLCA). The highly mutagenic cancer type *UCEC* harbored 11 of these shared IMMs (*DYM* R70Q, *NBN* S72Y, *PTEN* R142W, *RASA1* R427Q, *ROCK1* R1330Q, *SETD5* R882Q, *TOPORS* R347Q, *TCEA1* R153Q, *COPS2* S2967L, *ROCK1* D1014Y, *SASS6* R437Q, *ZNF283* R283I) (Table S9c).

Furthermore, 65.3% of the 156 IMMs shared by three or more patients were classified as driver missense mutations by CHASMplus (q<0.01). It is worth noting that while many shared IMMs were predicted to be driver missense mutations, the percentage IMMs predicted to be drivers was less than 1% of total IMMs in our study.

While we observed a limited number of shared IMMs, we reasoned that particular protein regions enriched for IMMs could present a therapeutic opportunity in certain cancer types. Using HotMaps 1D, we identified clusters of residues within protein regions having statistically significant enrichment of IMMs (q<0.1). These included *CIC* in low grade glioma (LGG) (7 IMMs between residues 202 and 260), *NFE2L2* (8 mutations, residues 24-81), *FGFR3* (22 IMMs, residues 216-249) (Figure 7D), *PIK3CA* (9 IMMs, residues 542-545) and *BIRC6* (5 IMMs, residues 440-480) in bladder cancer (BLCA), PTEN (28 IMMs, residues 95-173) and *CTNNB1* (22 IMMs, residues 32-41) in uterine corpus endometrial carcinoma (UCEC) (Table S9d).

We explored the relationship between mutation driver status predicted by CHASMplus, and IMM status using logistic regression. The log-odds of being an IMM was significantly decreased for drivers (β=-0.14, Wald test p=0.002), which is consistent with previous work suggesting that negative evolutionary selection eliminates MHC Class I immunogenic oncogenic mutations early in tumor development (57).

Finally, we considered whether a patient's IMM load was associated with changes in immune cell infiltrates as estimated from RNA sequencing of bulk cancer tissue. IMM load was significantly associated with increased total leukocyte fraction (β=0.77, Wald test p<2e-16) and with increased CD8+ T-cell fraction (β=3.4, Wald test p<2e-16).

These findings suggest a central role of IMMs in driving tumor immunoediting and may be informative for the interpretation of responses in the setting of immunotherapy.

Discussion

MHCnuggets provides a flexible open-source platform for MHC-peptide binding prediction that can handle common MHC Class I and Class II alleles, as well as rare alleles of both classes. The LSTM network architecture can handle peptide sequences of arbitrary length, without shortening or splitting. In addition, our neural network transfer learning protocols allow for parameter sharing among allele-specific, binding affinity -and HLAp-trained networks. When trained on binding affinity data, MHCnuggets achieves comparable performance to current methods. When trained on both binding affinity and HLAp data, we demonstrate significantly improved PPV_n on an independent HLAp test set, with respect to other methods that use both binding affinity and HLAp data. We attribute this improvement to both our choice of optimizing

PPV_n in our network training protocol and our implementation of transfer learning to integrate information from binding affinity and HLAp measurements.

We demonstrate improved scalability by comparing the runtime of MHCnuggets on 1 million peptides to comparable methods, and further by processing over 52 million expressed peptide-allele pairs across TCGA samples in under two hours. We identified 103,587 immunogenic missense mutations (IMMs) harbored by patients using 26 cancer types sequenced by the TCGA, based on transcriptional abundance and differential binding affinity compared to reference peptides. These results contrast with a previous report of neoantigens in TCGA patients in several respects. Rech *et al.* (50) applied a minimum expression threshold of 1 RNA sequencing read count, an IEDB-recommended combination of neoantigen predictors derived primarily from different versions of NetMHC, and IC50 threshold of 50nM to identify strong MHC binders. Their approach yielded 495,793 predicted Class I classically defined neoantigen peptides (each harboring a single immunogenic mutation) from 6,324 patients in 26 cancer types. As in our study, high variability in neoantigen burden across cancer types was observed. The striking difference between IMM and neoantigen burden in the two studies is likely due to differences in RNA expression threshold and the low false positive rate of MHCnuggets compared to IEDB-recommended tools.

Based on our conservative thresholds, IMMs were almost exclusively private to individual TCGA patients, with only 1,379 IMMs observed in more than one patient. Although more than 65% of IMMs shared by more than two patients were predicted to be driver mutations, the overall log odds of immunogenicity significantly decreased for predicted driver mutations, indicating immunogenicity might shape the driver mutation landscape. Patient IMM counts were also significantly associated with increase in total leukocyte fraction and fraction of CD8+ T-cells, suggesting that they may be relevant to immune system response to cancer.

This work has several limitations. First, our analyses are limited to missense mutations, and while these are very numerous, there is substantial evidence that somatic gene fusions, frameshift indels, splice variants etc. in tumors may also generate neoantigens. Next, recent work suggests that peptidal context, such as flanking sequence, its source protein and the expression level of the source protein, is informative for MHC ligand prediction (26,33). This type of information is currently only available for a limited number of HLAp data sets, which were unavailable to us for training purposes. As more well-characterized HLAp datasets become available, we will extend MHCnuggets to include these features. We did not address T-cell receptor (TCR) binding to bound peptide-MHC complexes or T-cell activation upon complex binding. While we are actively pursuing this more complex modeling problem, we believe that improved prediction of peptide binding to MHC is also therapeutically relevant (26). Finally, we are unable to directly compare performance to the MHC Class II prediction methods from the NetMHC group, except for self-reported auROC. While we are not able to do a rigorous comparison of MHCnuggets Class II prediction, our benchmark comparisons suggested that MHCnuggets was competitive with NetMHCII2.3 and that MHCnuggets Class II rare allele performance was competitive with NetMHCIIpan3.2. Generally rare allele performance estimated for each allele, regardless of MHC Class or performance metric was variable among individual alleles for both MHCnuggets and NetMHCIIpan3.2, suggesting that further work in this area is warranted.

In summary, we present MHCnuggets, an open source software package for MHC ligand prediction that improves on performance of previous methods with respect to positive predictive value by leveraging transfer learning to integrate binding affinity and HLAp data. In contrast to previous methods, it handles both MHC Class I and Class II ligand prediction and both common and rare HLA alleles, within a single framework. The utility of MHCnuggets is demonstrated with a basic pipeline for large-scale cancer patient sequencing data from TCGA, which analyzed mutation immunogenicity, shared IMMs and the relationship between mutation immunogenicity, driver potential and immune infiltrates.

Financial support

This work was supported in part by the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up to Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Commonwealth Foundation, U.S. NIH (grants CA121113, CA006973, and CA180950), the V Foundation and LUNGevity. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research.

Disclosure of potential conflicts of interest

V.A receives research funding from Bristol-Myers Squibb. V.E.V. is a founder of Personal Genome Diagnostics, a member of its Scientific Advisory Board and Board of Directors, and owns Personal Genome Diagnostics stock, which are subject to certain restrictions under university policy. V.E.V. is an advisor to Takeda Pharmaceuticals. Within the last five years, V.E.V. has been an advisor to Daiichi Sankyo, Janssen Diagnostics, and Ignyta. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies.

Acknowledgments

Part of this research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC).

Figures:

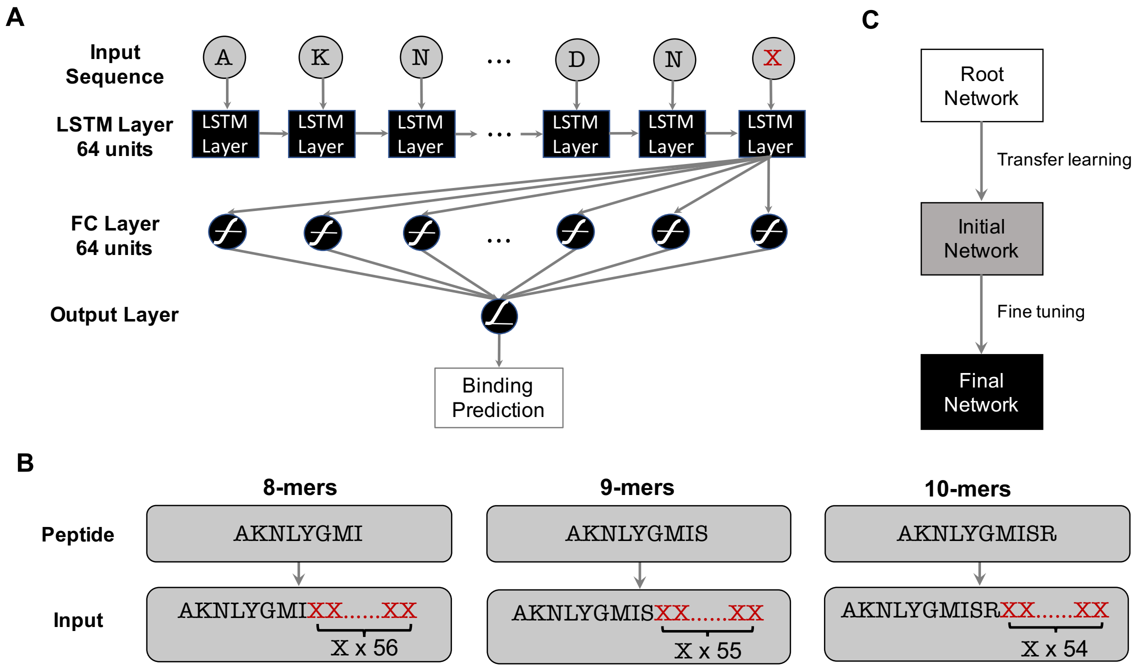


Figure 1. A) MHCnuggets architecture. B) Input encoding scheme for peptides with variable lengths. MHCnuggets accepts peptides with length up to 64 amino acid residues. **C) Transfer learning protocol for parameter sharing among alleles.** The base network is selected to be most abundantly represented allele in the training set.

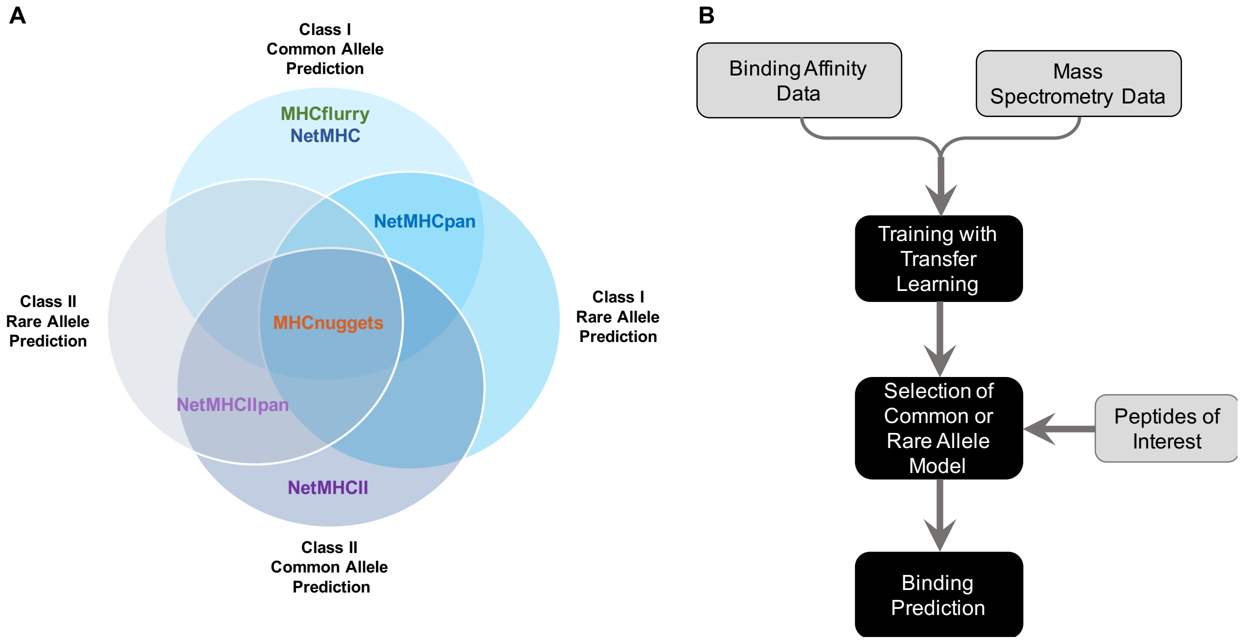


Figure 2. MHCnuggets. A) Venn diagram representation of the MHC-peptide binding prediction functions of MHCnuggets and several other currently available tools. B) Training and MHC allele model selection scheme for MHCnugget**s.**

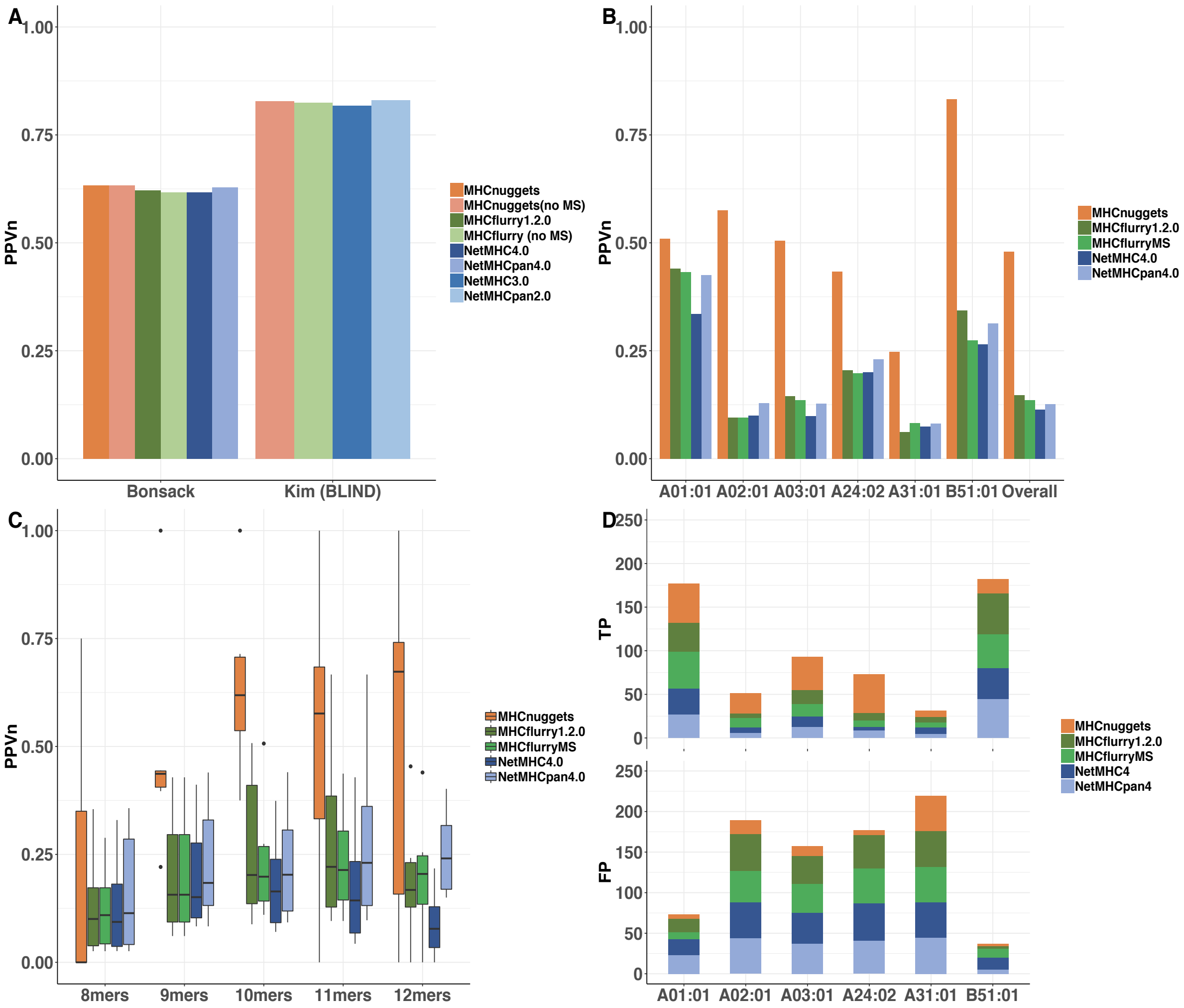


Figure 3. MHC Class I benchmark comparisons. A) PPV_n for MHC Class I allele-specific prediction on binding affinity test sets from Bonsack et al. (7 alleles) and Kim *et al*. (53 alleles) B) PPV_n for MHC Class I allele-specific prediction on HLAp BST data set (Bassani-Sternberg et al. and Trolle *et al*.), stratified by allele (6 alleles). C) PPV_n for MHC Class I allele-specific prediction on HLAp BST data set (from B) stratified by peptide sequence length. D) True and false positives for each method on the top 50 ranked peptides from the HLAp BST data set. PPV_n = positive predictive value on the top *n* ranked peptides, where *n* is the number of true binders. TP=true positives. FP=false positives.

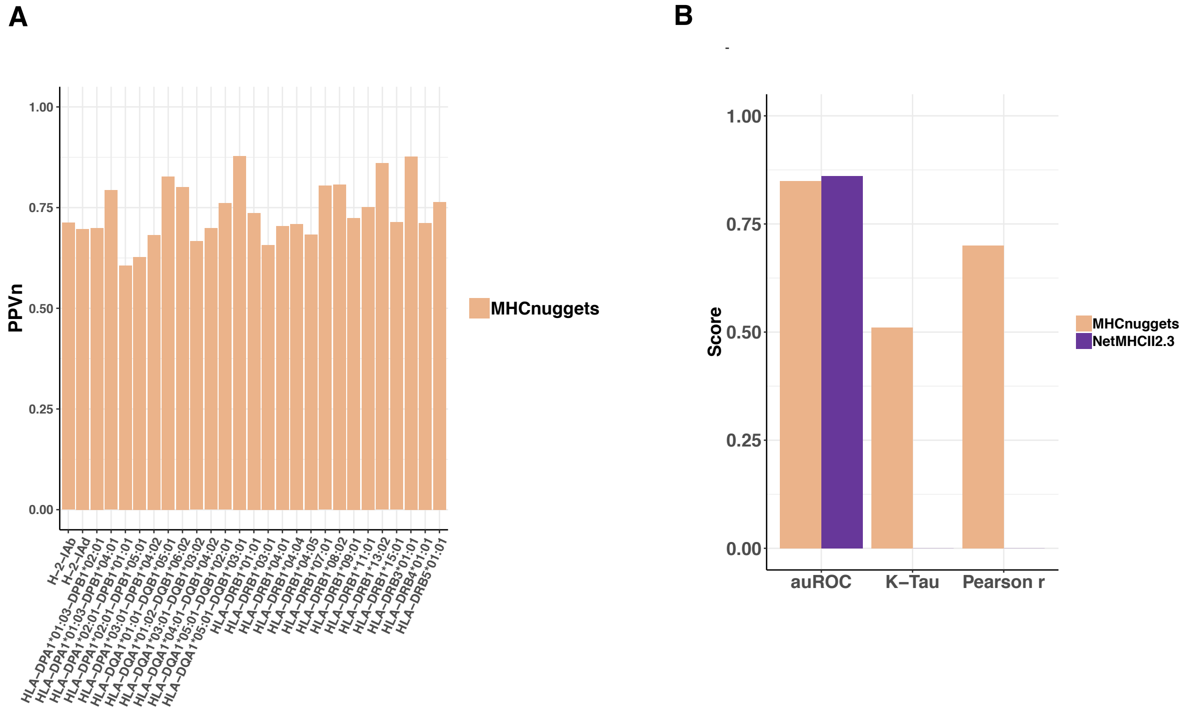


Figure 4. MHC Class II benchmark comparisons. A) PPV$_n$ for MHC Class II allele-specific prediction on binding affinity test set from Jensen *et al.* (27 alleles, stratified by allele). B) auROC, K-Tau, Pearson r scores for MHC Class II alleles from five-fold cross-validation. NetMHCII2.3 performance is from their self-reported auROC. auROC= area under the receiving operator characteristic curve. K-Tau = Kendall's *tau* correlation. PPV$_n$ = positive predictive value on the top *n* ranked peptides, where *n* is the number of true binders.

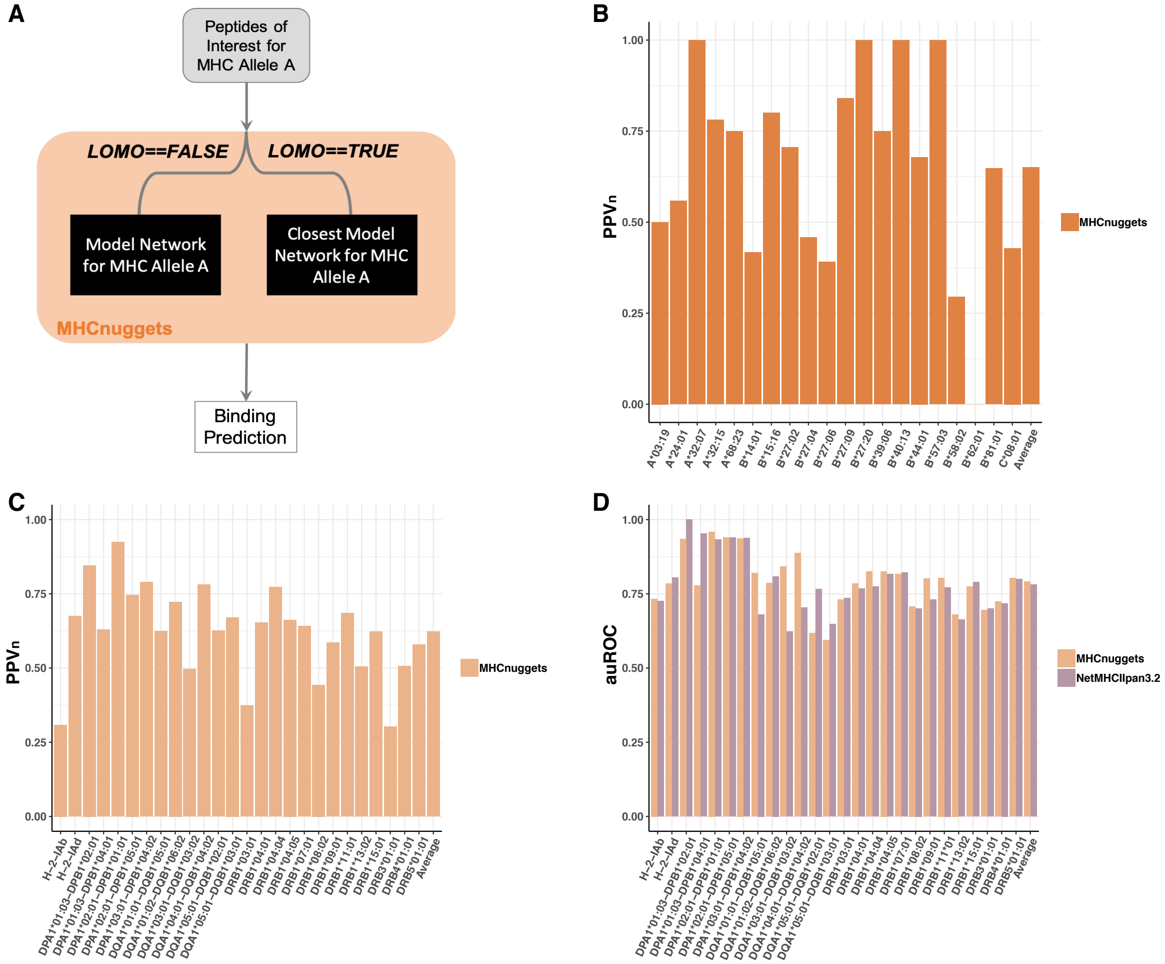


Figure 5. MHC Class I and II benchmark comparisons to estimate rare allele performance. A) Schematic representation of leave one molecule out (LOMO) testing. B) PPV_n for MHC Class I rare allele prediction on IEDB pseudo-rare alleles binding affinity test set (20 alleles, stratified by allele). C) PPV_n for MHC Class II rare allele prediction on binding affinity test set from Jensen et al. (27 alleles, stratified by allele). D) auROC for MHC Class II rare allele prediction on LOMO binding affinity test set from Jensen *et al*. (27 alleles, stratified by allele). NetMHCIIpan3.2 results are from their self-reported auROC. auROC = area under the receiving operator characteristic curve. PPV_n = positive predictive value on the top *n* ranked peptides, where *n* is the number of true binders.

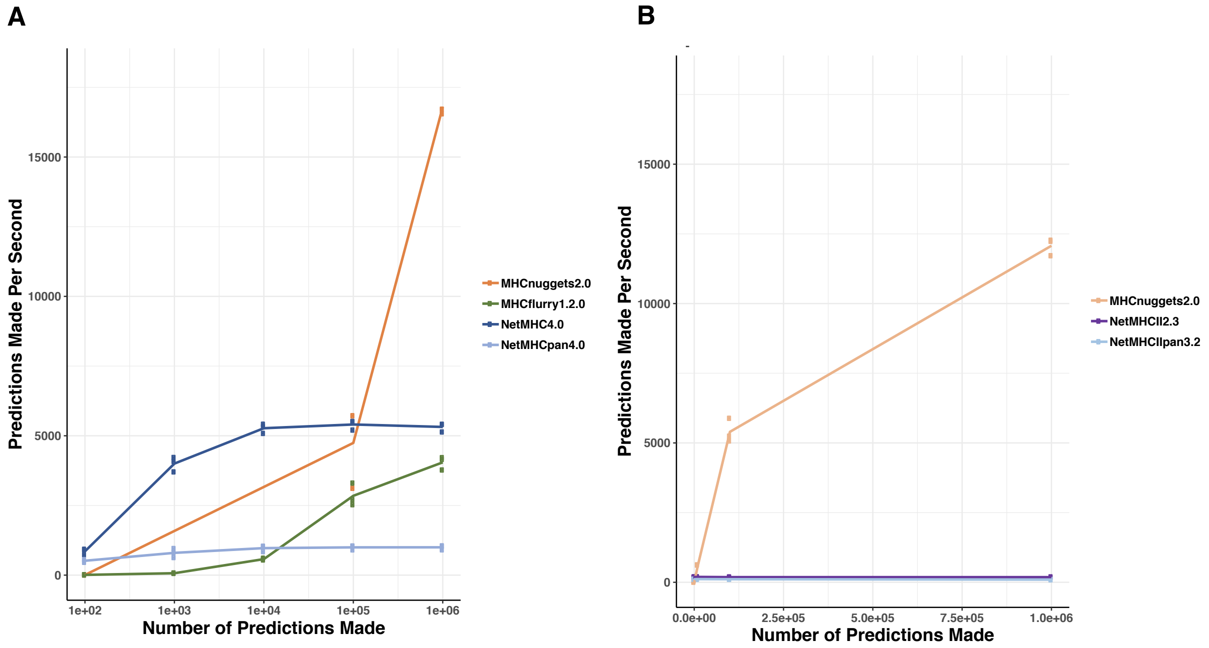


Figure 6. Timing and scalability. Runtime benchmark of most recent version of tested methods over a range of inputs (up to 1 million peptides). A) MHC Class I prediction. B) MHC Class II prediction

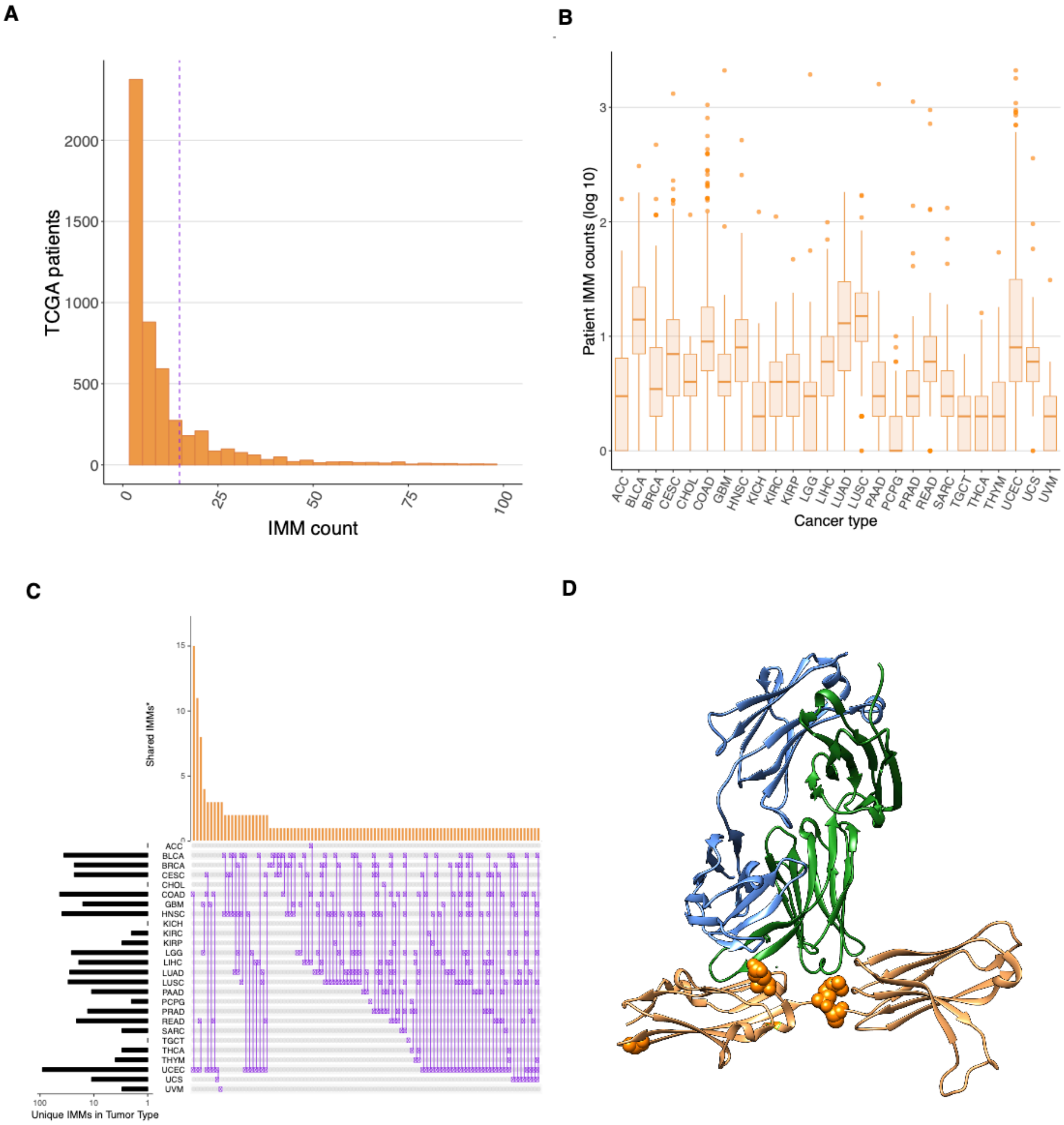


Figure 7. MHC Class I IMMs in TCGA patients. A). Number of candidate immunogenic missense mutations (IMMs) identified in 6,613 TCGA patients. Dotted line = mean IMMs per patient (14.9). B) Number of candidate IMMs by cancer type. C) IMMs shared by three or more patients and the cancer types in which they occurred. Each row represents a cancer type and each column illustrates the overlap of IMMs seen in a single cancer type or multiple cancer types. For example, the first column shows the number of IMMs shared among patients with colorectal

adenocarcinoma (COAD) and uterine corpus endometrial carcinoma (UCEC). Bars to the left show the total number of unique IMMs in each cancer type. *Bar heights are count of unique shared IMMs, not total number of patients in which the IMM was observed. Cancer type abbreviations are in Supplementary Methods. Image generated with UpSetR (58). D) Fibroblast growth factor receptor (*FGFR3*) IMM hot region identified by HotMAPs in bladder cancer (BLCA). IMMs shown and number of BLCA patients with the IMM: p.E216K (1), p.G235D (1) p.R248C (2) and p.S249C (18). Except for p.G235D, these IMMs are proximal to the interface of FGFR3 protein and the light and heavy chains of an antibody fragment designed for therapeutic application in bladder cancer (PDB ID: 3GRW) (59).

References

1. Parmiani G, Castelli C, Dalerba P, Mortarini R, Rivoltini L, Marincola FM*, et al.* Cancer Immunotherapy With Peptide-Based Vaccines: What Have We Achieved? Where Are We Going? JNCI: Journal of the National Cancer Institute **2002**;94:805-
2. Lu Y-C, Robbins PF. Cancer immunotherapy targeting neoantigens. Seminars in Immunology **2016**;28:22-7
3. Reinherz EL. αβ TCR-Mediated Recognition: Relevance to Tumor-Antigen Discovery and Cancer Immunotherapy. Cancer Immunology Research **2015**;3:305-12
4. Wang R-F, Wang HY. Immune targets and neoantigens for cancer immunotherapy and precision medicine. Cell Res **2017**;27:11-37
5. Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J*, et al.* Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non–Small Cell Lung Cancer. Cancer Discovery **2017**
6. Yarchoan M, Johnson BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nature reviews Cancer **2017**;17:209-22
7. Lundegaard C, Lund O, Buus S, Nielsen M. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology **2010**;130:309-18
8. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics **2016**;32:511-7
9. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC Bioinformatics **2014**;15:241-
10. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. BMC Bioinforma **2009**;10
11. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O*, et al.* Automated benchmarking of peptide-MHC class I binding predictions. Bioinformatics **2015**;31:2174-
12. Bonsack M, Hoppe S, Winter J, Tichy D, Zeller C, Kupper MD*, et al.* Performance Evaluation of MHC Class-I Binding Prediction Tools Based on an Experimentally Validated MHC-Peptide Binding Data Set. Cancer Immunol Res **2019**;7:719-36
13. Gfeller D, Bassani-Sternberg M, Schmidt J, Luescher IF. Current tools for predicting cancer-specific T cell immunity. OncoImmunology **2016**;5:e1177691-e
14. Liu XS, Mardis ER. Applications of Immunogenomics to Cancer. Cell **2017**;168:600-12
15. Hachinski V. The treatment of low-grade glioma. Arch Neurol **1989**;46:1239
16. Editorial NB. The problem with neoantigen prediction. Nature Biotechnology **2017**;35:97
17. Wieczorek M, Abualrous ET, Sticht J, Alvaro-Benito M, Stolzenberg S, Noe F*, et al.* Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Front Immunol **2017**;8:292
18. Lu YC, Robbins PF. Targeting neoantigens for cancer immunotherapy. Int Immunol **2016**;28:365-70

19. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. The Journal of Immunology **2017**

20. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. Immunogenetics **2013**;65:711-24

21. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Research **2008**;36:509-12

22. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S*, et al.* NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. PLOS ONE **2007**;2:1-10

23. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst **2018**;7:129-32 e4

24. Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. Curr Opin Immunol **2016**;41:9-17

25. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science **2015**;348:69-74

26. Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A*, et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nat Biotechnol **2018**

27. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics **2015**;14:658-73

28. Berlin C, Kowalewski DJ, Schuster H, Mirza N, Walz S, Handel M*, et al.* Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. Leukemia **2015**;29:647-59

29. Kalaora S, Barnea E, Merhavi-Shoham E, Qutob N, Teer JK, Shimony N*, et al.* Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. Oncotarget **2016**;7:5110-7

30. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO*, et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol **2017**;13:e1005725

31. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation **1997**;9:1735-80

32. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. 2018. Springer. p 270-9.

33. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J*, et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity **2017**;46:315-26

34. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S*, et al.* IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res **2015**;43:D413-22

35. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A*, et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. Immunome Res **2008**;4:2
36. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics **2011**;63:325-35
37. Abadi Mn, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C*, et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. CoRR **2016**;abs/1603.04467
38. Chollet F, others. Keras. GitHub; 2015.
39. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER*, et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med **2016**;8:11
40. Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K*, et al.* Population-level distribution and putative immunogenicity of cancer neoepitopes. BMC Cancer **2018**;18:414
41. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z*, et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology **2018**;154:394-406
42. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR*, et al.* The immune epitope database (IEDB) 3.0. Nucleic Acids Research **2015**;43:D405-D
43. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A*, et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell **2018**;173:371-85 e18
44. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C*, et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst **2018**;6:271-81 e7
45. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH*, et al.* The Immune Landscape of Cancer. Immunity **2018**;48:812-30 e14
46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **2011**;12:323
47. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics **2014**;30:3310-6
48. UniProt C. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res **2014**;42:D191-8
49. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP*, et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. J Exp Med **2014**;211:2231-48
50. Rech AJ, Balli D, Mantero A, Ishwaran H, Nathanson KL, Stanger BZ*, et al.* Tumor Immunity and Survival as a Function of Alternative Neopeptides in Human Cancer. Cancer Immunol Res **2018**

51. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M*, et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res **2016**;76:3719-31

52. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **1995**;57:289-300

53. Tokheim C, Karchin R. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. Cell Syst **2019**

54. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y*, et al.* Robust enumeration of cell subsets from tissue expression profiles. Nat Methods **2015**;12:453-7

55. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA oncogene in human cancers. Br J Cancer **2006**;94:455-9

56. Tomlinson DC, Hurst CD, Knowles MA. Knockdown by shRNA identifies S249C mutant FGFR3 as a potential therapeutic target in bladder cancer. Oncogene **2007**;26:5889-99

57. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB*, et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell **2017**;171:1272-83 e15

58. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics **2017**;33:2938-40

59. Qing J, Du X, Chen Y, Chan P, Li H, Wu P*, et al.* Antibody-based targeting of FGFR3 in bladder carcinoma and t(4;14)-positive multiple myeloma in mice. J Clin Invest **2009**;119:1216-29

60. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S*, et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci **2003**;12:1007-17

61. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR **2014**;abs/1412.6980

62. Gal Y. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. arXiv:151205287 **2015**

63. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR*, et al.* The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res **2019**;47:D339-D43

64. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O*, et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics **2009**;61:1-13

65. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol **2018**;1711:243-59

Supplementary Information

Methods

Implementation

Transformation of peptide binding affinities. Predicted binding affinity can be transformed into a range of values well-suited for neural network learning by selecting a logarithmic base to match the weakest binding affinity of interest (60). For most benchmarks in this work, we used the standard upper limit of 50,000 nM, so that predicted binding affinity was $y = max(0,1 - log_{50k}(IC50))$. For the Bonsack et al. dataset (12), the upper limit was changed to 100,000nM because in their experiments, as described in O'Donnell et al. (23), binders were defined as peptides with IC50<100,000nM. As binding affinity was determined based on in vitro HLA binding-competition vs. a known strong binder (reported IC50 <50nM) experimental IC50 values were in µM range.

Selection of final network weights. To minimize overfitting, network training was stopped after 100 epochs but if the best PPV$_n$ was reached earlier, network weights from that earlier epoch were used in the final network. Notably, while we chose to optimize the networks on PPV$_n$, an alternative approach could optimize on auROC, Kendall's *tau* or Pearson's *r* correlation. For the two alleles in IEDB with the most training examples in their respective class, HLA-A*02:01 for Class I and HLA-DRB2*01 for Class II, training was stopped after 200 epochs.

Network training. Mean-squared error loss L_{MSE} was used to train networks with continuous-valued binding affinity data and binary cross-entropy loss L_{BCE} for binary HLAp data. For a dataset with *n* samples,

$$L_{MSE}(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

$$L_{BCE}(\hat{y}, y) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})$$

All training used backpropagation with the Adam optimizer (61) and learning rate of 0.001. Regularization was performed with dropout and recurrent dropout (62) probabilities of 0.2. The number of hidden units, dropout rate, and number of training epochs was estimated by three-fold cross-validation on MHC Class I A*02:01, a common allele with a large number of experimentally characterized binding peptides.

One-hot encoding: Peptides were represented to the network as a series of amino acids; each amino acid was represented as a 21-dimensional smoothed, one-hot encoded vector (0.9 and 0.005 replace 1 and 0, respectively).

Transfer Learning Protocol for binding affinity data only. We used transfer learning to improve network learning for MHC alleles with limited characterized peptides available for training. We first trained base allele-specific networks for Class I and Class II, using alleles with the most training examples in IEDB (HLA-A*02:01 for Class I and HLA-DRB2*01 for Class II). For all other alleles, the final weights of the base network for its respective class were used to initialize network training, and then an allele-specific network was trained for each allele. Next, we assessed prediction performance of each allele-specific network on the training examples for each of the alleles. For each allele, if the network that performed best was not the HLA-A*02:01 network (for Class I alleles) or HLA-DRB1*01:01 network (for Class II alleles), we did a second round of training, with the best performing network's weights used in the initialization step.

Transfer Learning Protocol for binding affinity and HLAp data. To integrate HLAp data into the Class I networks we initially trained each network with binding affinity data as described above, transferred the final weights to a new network, and then continued training with the HLAp data as positive examples augmented with random peptide decoys as negative examples.

Dataset collection and curation

Table S1. Data sources used in training and benchmarking. BA=binding affinity. HLAp=peptide elution/mass spectrometry. LOMO=leave one molecule out cross-validation. Only alleles with >30 characterized peptides were included. IEDB=curated version 2018. Common allele=>30 peptides with characterized binding information, rare allele= <30 characterized peptides, mono-allelic = engineered cells that express a single MHC allele, multi-allelic = cells that express multiple MHC alleles. Abelin et al. (33), Bonsack et al. (12), Kim et al. (9) Bassani-Sternberg et al. (30) Trolle et al. (11), Jensen et al. (41), IEDB (63).

MHC Class	Common or rare alleles	**Training Sets**		**Benchmarks**	
		Datasets	**Description**	**Datasets**	**Description**
Class I	Common	*IEDB*	BA, 241,553 peptides for 217 alleles	*Bonsack et al*	BA, Tested on 475 peptides for 7 alleles
		Abelin et al	HLAp, 23,651 peptides for 16 alleles, mono-allelic	*Kim et al*	BA, Trained on 53 alleles (BD2009), tested on 53 alleles (BLIND)
				BST	HLAp, 29,501 hits for 6 alleles .from Bassani-Sternberg 2017 and Trolle et al. plus random peptide decoys.
				Bassani-Sternberg et al 2017	HLAp, 22,598 hits for 26 alleles, multi-allelic Included in BST.
				Trolle et al	HLAp, 15,524 hits for 5 alleles, mono-allelic Included in BST.
				Random peptide decoys from human proteome	29471500 random decoy peptides generated from the human proteome (courtesy of Sisi Sarkizova and Cathy Wu) Included in BST.
Class II	Common	*IEDB*	BA, 96,211 peptides for 135 alleles	*Jensen et al*	BA, Five-fold cross-validation on 27 alleles.
Class I	Rare	NA	NA	*IEDB Pseudo-Rare Alleles*	BA, LOMO on 20 alleles, alleles with training samples between 30 and 100
Class II	Rare	NA	NA	*Jensen et al*	BA, LOMO on 27 alleles

Data sources for network training and testing, TCGA somatic mutations, TCGA tumor gene expression and haplotype calling are shown in Table S1 and Table S2. A curated version of the IEDB database 2018 (63) and the sixteen Class I mono-allelic B-cell line immunopeptidomes (33) was provided by Tim O'Donnell (https://data.mendeley.com/datasets/8pz43nvvxh/2), binding affinity assays of HPV-derived peptides were provided by Maria Bonsack and Angelika Riemer (12), BST = immunopeptidomes from six cell lines with multi-allelic MHCs (26 MHC Class I alleles) (30) and from soluble HLA(sHLA)-transfected HeLa cells separated by allele (4 MHC Class I alleles) (11). Decoy random peptides sampled from the human proteome were generously provided by Cathy Wu (33).

Kim et al: This benchmark contained 53 MHC Class I alleles and 137,654 IC50 measurements published prior to 2009 (training set) and 53 unique MHC Class I alleles with 26,888 IC50 measurements, published from 2009-2013 (test set). Three alleles (HLA-B*27:03, HLA-B*38:01, HLA-B*08:03) did not contain sufficient training data, and two alleles (HLA-A*46:01, HLA-B*27:03) did not contain any peptides defined as binders in this work (IC50<500nM). Therefore, a total of four alleles (HLA-A*46:01,HLA-B*27:03, HLA-B*38:01, and HLA-B*08:03) were dropped from the analysis. All peptides in this benchmark set consisted of 8-11 amino acid residues.

Bonsack et al. This dataset contains 475 synthetic peptides derived from model protein sequences HPV16 E6 and E7 tested for binding to 7 alleles (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*11:01, HLA-A*24:02, HLA-B*07:02 and HLA-B*15:01). Each peptide was tested in competition-based cellular binding assays with a known high-affinity fluorescein-labeled reference peptide. EBV-transformed B-lymphoblastic cells were stripped of their naturally-bound peptides and mixed with serially diluted test peptides and 150 nM of reference peptide. Each synthetic peptide was tested at 8 different concentrations ranging from 780 nM to 100,000 nM. Mixture fluorescence at each synthetic peptide concentration was measured with flow cytometry, and a non-linear regression analysis was used to find the test peptide concentration that inhibited 50% of the reference peptide binding (IC50). Peptides were classified as binders (IC50 <= 100,000 nM) or nonbinders (IC50 > 100,000 nM). Peptides in this independent benchmark set do not have IEDB entries.

Bassani-Sternberg et al. 2017 (30) This dataset contains 22,598 unique peptides eluted from 6 cell lines with multi-allelic MHCs. Out of the total 6 cell lines, a total of 26 alleles were reported. For each multi-allelic cell line, peptide/MHC pairs were found through deconvolution, following the protocol described by (33), with the difference that we used MHCnuggets rather than NetMHCpan2.8 (64) to predict IC50 values for each peptide-MHC pair. For each cell line, each peptide was initially assigned as a binder to all expressed alleles. Then, for each allele, we filtered out any peptide predicted to bind with IC50>1000nM to that allele, and with

IC50<150nM to any other allele. Peptides found for 6 alleles (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*24:02, HLA-A*31:01, HLA-B*51:01), were selected for allele-specific prediction testing. Trained networks were available for these alleles from all the methods that we compared.

Trolle et al. This dataset contains 15,524 unique peptides eluted from soluble HLA(sHLA) transfected HeLa cells, a process that allowed for separating binding peptides to a single MHC allele. This dataset reports peptides for 5 MHC alleles. Peptides found for 4 alleles (HLA-A*01:01, HLA-A*02:01, HLA-A24*02, HLA-B*51:01) were selected for testing. Peptide lengths in this dataset range from 8-15 amino acid residues.

BST. This benchmark consists of 29,501 HLAp hits for 6 alleles, from Bassani-Sternberg et al. 2017 and Trolle et al. plus 29471500 random decoy peptides and was used as an independent test set of HLAp data for neural networks trained on Abelin et al. in a previous work (33).

Jensen et al. This benchmark was designed to assess both allele-specific and rare MHC Class II binding affinity predictors. Allele-specific prediction was tested with a five-fold cross validation experiment on peptides found in IEDB in 2016 but not 2013. Rare allele predictions were tested with the LOMO protocol.

IEDB Class I rare alleles. This dataset was designed to apply the LOMO protocol to Class I alleles. It included 20 "pseudo-rare" alleles with 30-100 binding affinity peptide measurements in IEDB.

Performance metrics

We calculated positive predictive value with respect to the top-ranked *n* peptides, where *n* is the number of true binders in the ranked list, denoted as PPV_n.
PPV = NTP / (NTP+NFP), where NTP=number of true positives and NFP=number of false positives. We calculated PPV with respect to the top-ranked n peptides, where n is the number of true binders in the ranked list, denoted as PPV_n. For the BST benchmark, we also calculated PPV over the top 50 and 500 ranked peptides.

TCGA analysis pipeline

MC3 mutation filtering: MC3 TCGA somatic mutation calls were filtered with the same procedure used in the TCGA PanCan Atlas Drivers Analysis Working Group paper, including exclusion of highly mutated (hypermutator) samples. A hypermutator was defined as a sample with a mutation count exceeding Tukey’s outlier condition, of >1.5 times the interquartile range above the third quartile (3Q + 1.5*IQR) in its cancer type and number of mutations in a sample>1000 (43).

Cancer types in the TCGA are abbreviated as follows: Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG), Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Pancreatic adenocarcinoma (PAAD), Prostate adenocarcinoma (PRAD), Rectum adenocarcinoma (READ), Sarcoma (SARC), Thymoma (THYM), Thyroid carcinoma (THCA), Uterine Carcinosarcoma (UCS), Uterine Corpus Endometrial Carcinoma (UCEC), Uveal Melanoma (UVM).

Regression models: We applied two univariate Poisson regression models. In the first model, each patient's immunogenic missense mutation load was the response variable and X_1 was the total leukocyte fraction. The fitted coefficient $\beta_1 = 0.77$ (p<2e-16, Wald test) indicated that increased IMM load was associated with increased leukocyte fraction in a patient's cancer. In a second model, X_1 was the proportion of CD8+ T cells inferred by CIBERSORT (65). The fitted coefficient $\beta_1 = 3.4$ (p<2e-16, Wald test) indicated that increased IMM load was associated with increased tumor-infiltrating CD8+ T cells. Total lymphocyte and (Aggregate3) CD8+ T cell fractions were estimated in Thorsson *et al.* (45).

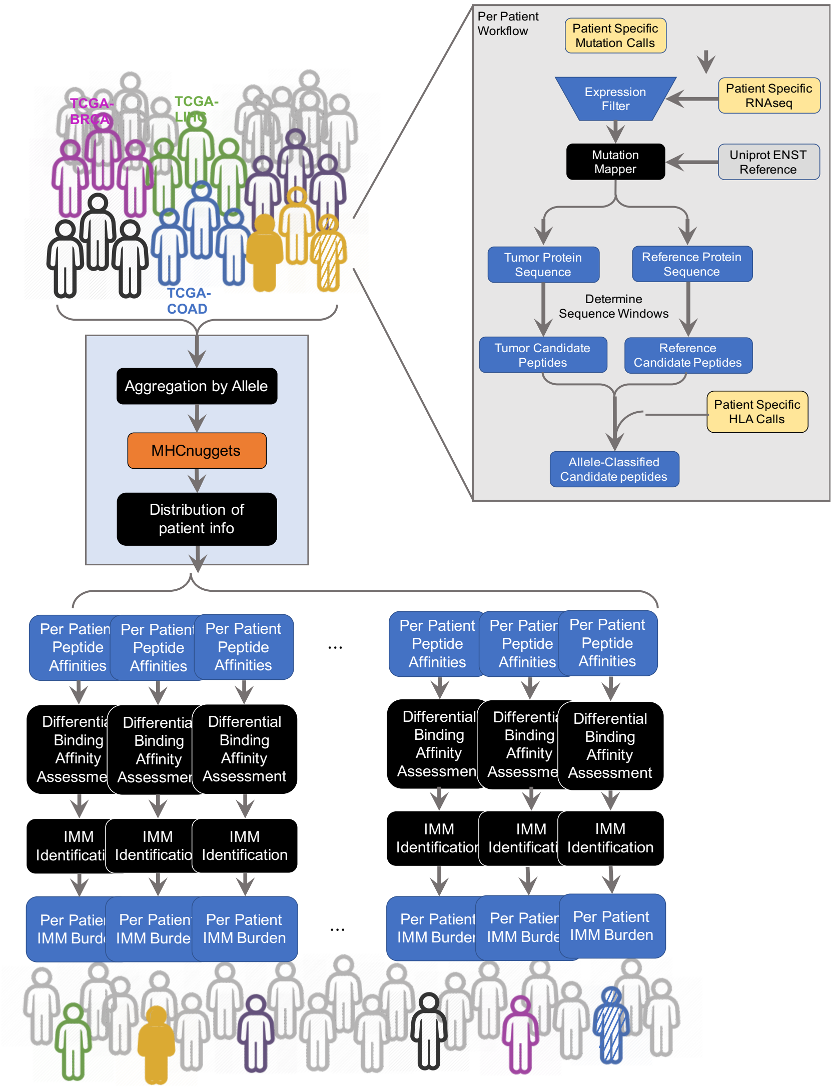
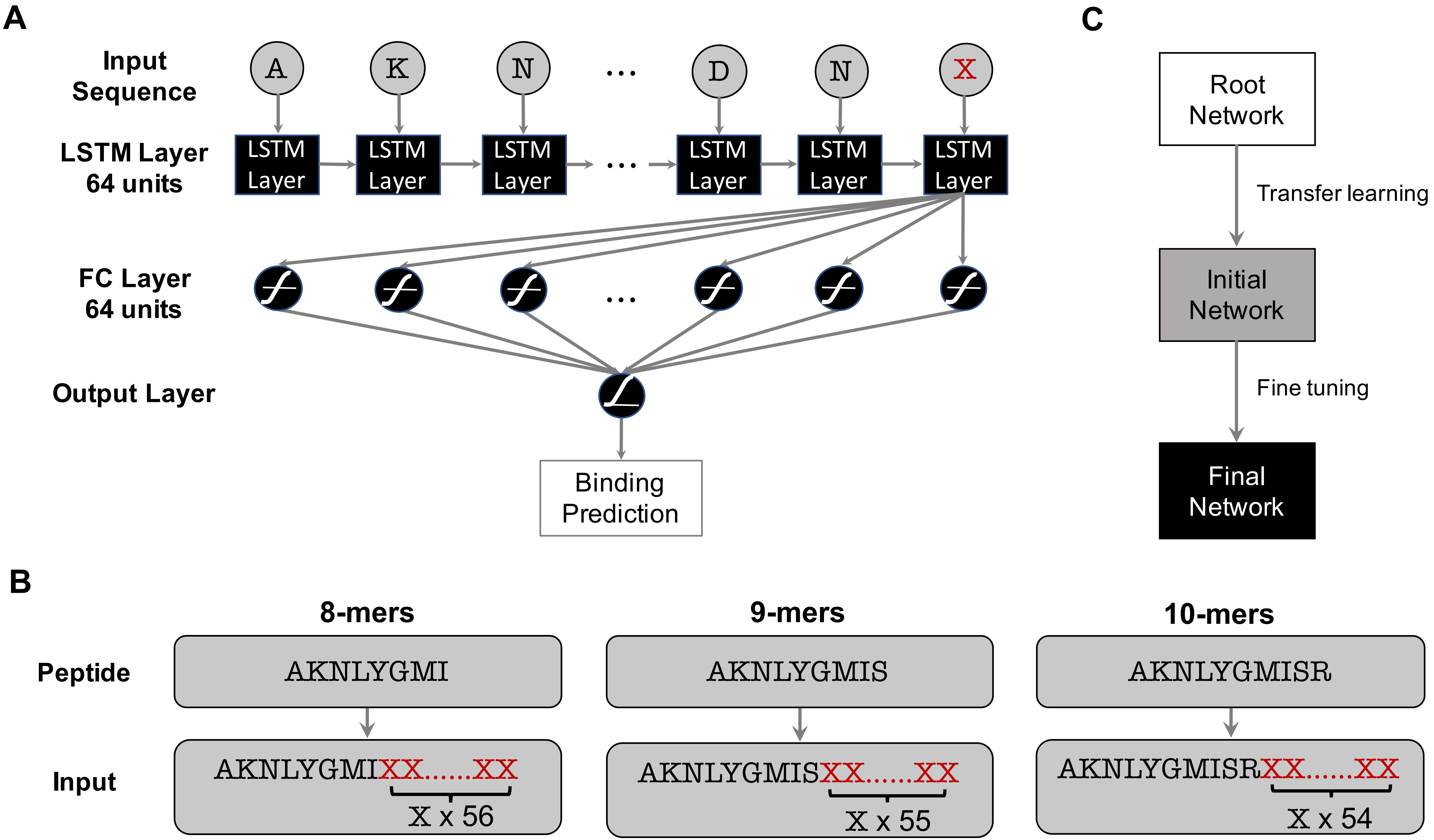
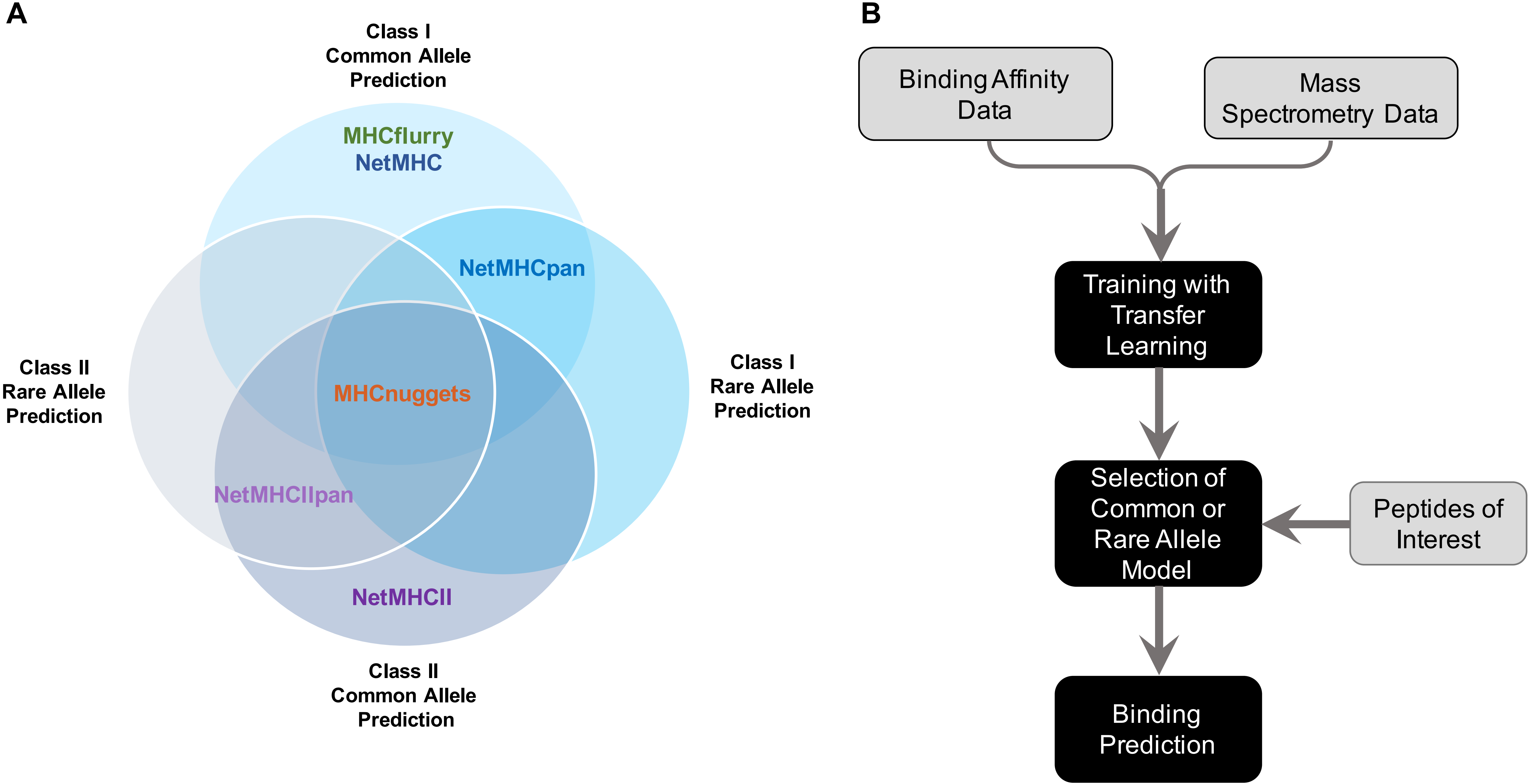
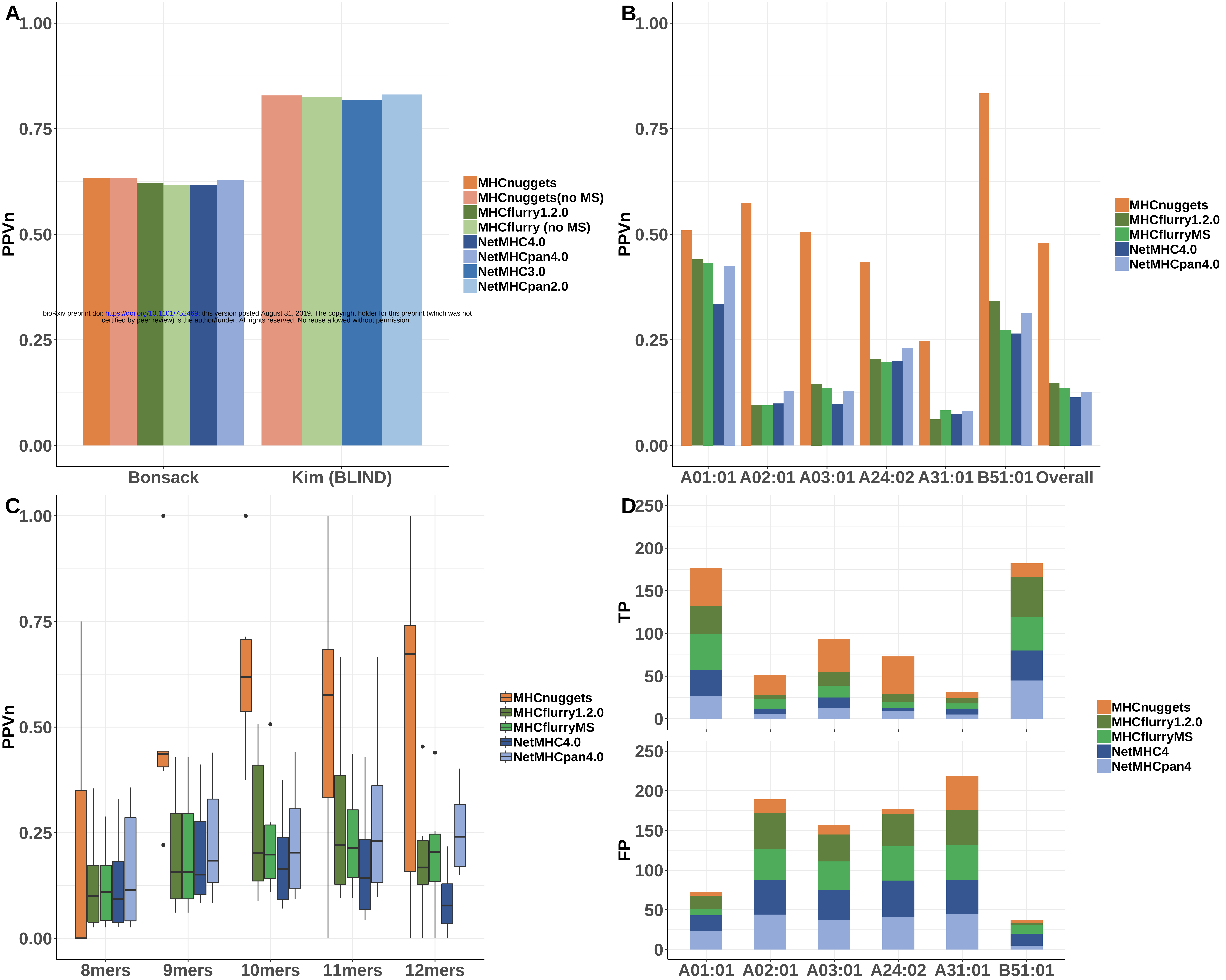
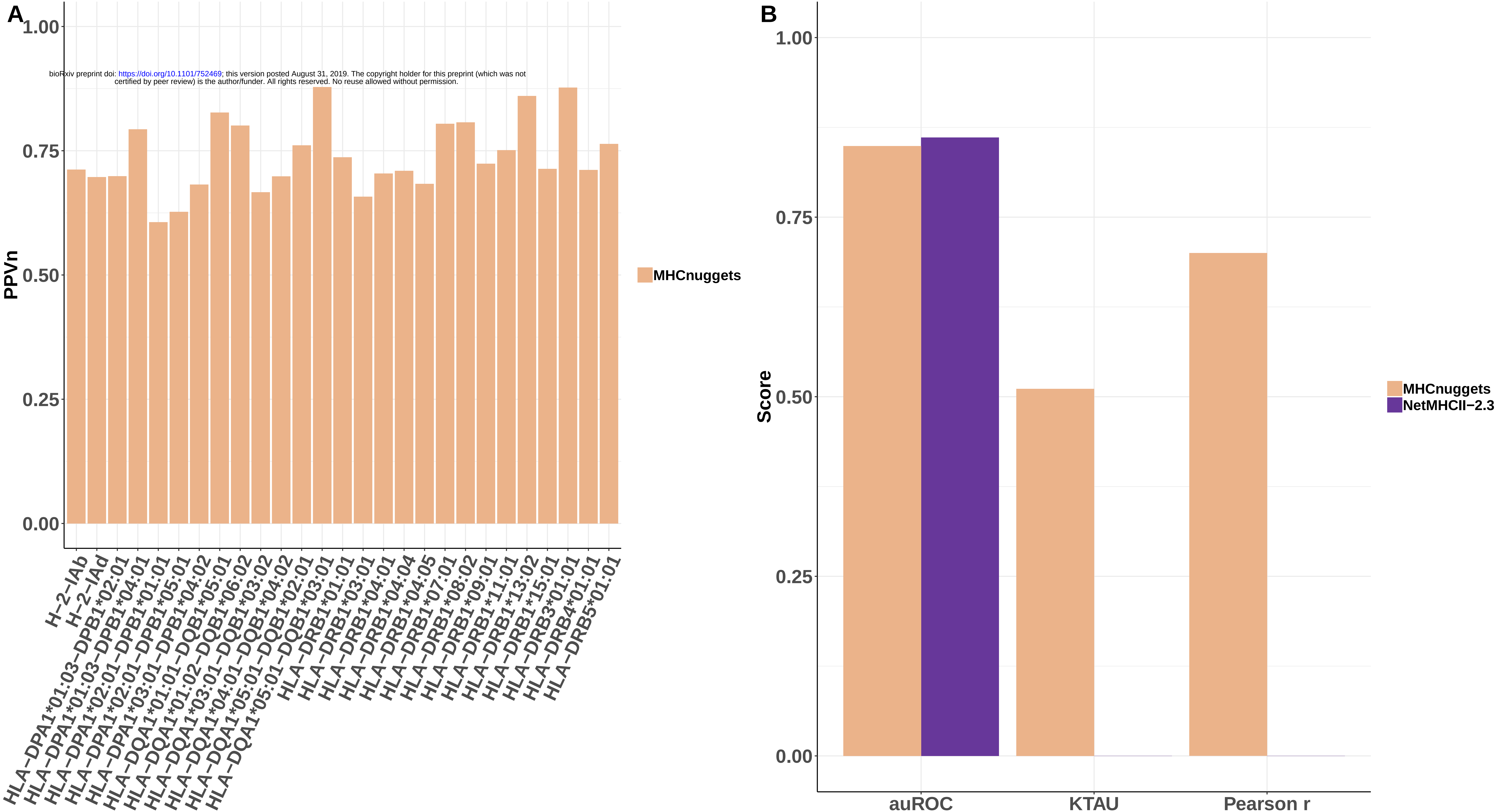

TCGA-BRCA
TCGA-LIHC
TCGA-COAD
Aggregation by Allele
MHCnuggets
Distribution of patient info
Per Patient Workflow
Patient Specific Mutation Calls
Expression Filter
Patient Specific RNAseq
Mutation Mapper
Uniprot ENST Reference
Tumor Protein Sequence
Reference Protein Sequence
Determine Sequence Windows
Tumor Candidate Peptides
Reference Candidate Peptides
Patient Specific HLA Calls
Allele-Classified Candidate peptides
Per Patient Peptide Affinities
...
Differential Binding Affinity Assessment
IMM Identification
Per Patient IMM Burden
...

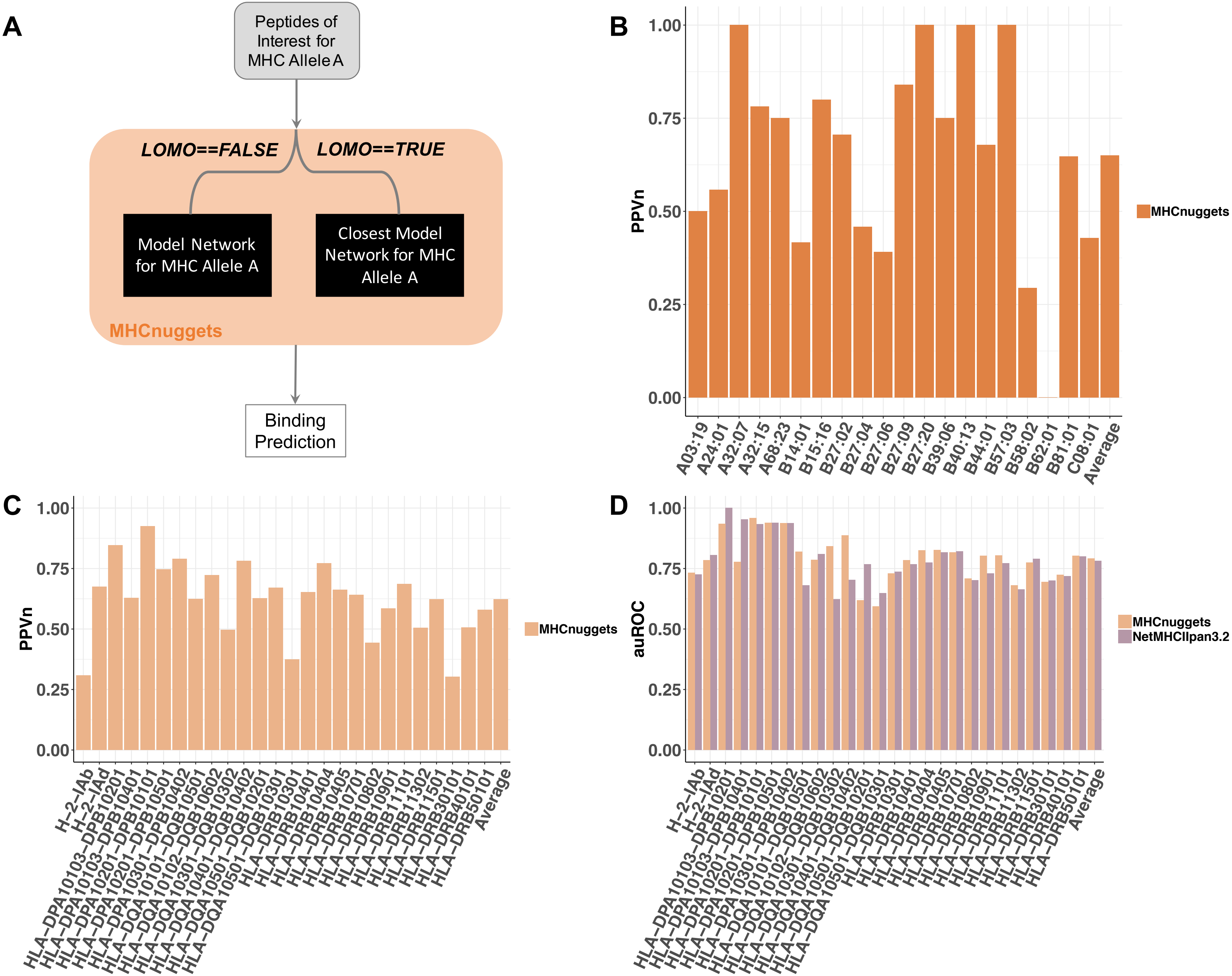
Figure S1. Flow chart describing the neoantigen prediction pipeline applied to TCGA. A. Per patient workflow. TCGA mutation calls were filtered by transcript expression for each patient. Mutations were mapped to reference transcripts and protein sequences. Peptides of length 8-11 were generated based upon reference and mutated sequences. Candidate peptides for each mutation were selected by differential binding affinity to up to six possible Class I alleles from each patient (Methods). B. Hourglass data processing of TCGA samples. Peptides were aggregated by allele, and differential agretopic index based on MHCnuggets predicted binding affinities was calculated across all patients for each allele. Peptides that passed all filters were considered candidate neoantigens and were re-assigned to the originating patient. C. Somatic missense mutations included in candidate neoantigen peptides were considered to be IMMs, and IMM load was computed for each patient and in aggregate for each cancer type. IMM=immunogenic missense mutation.

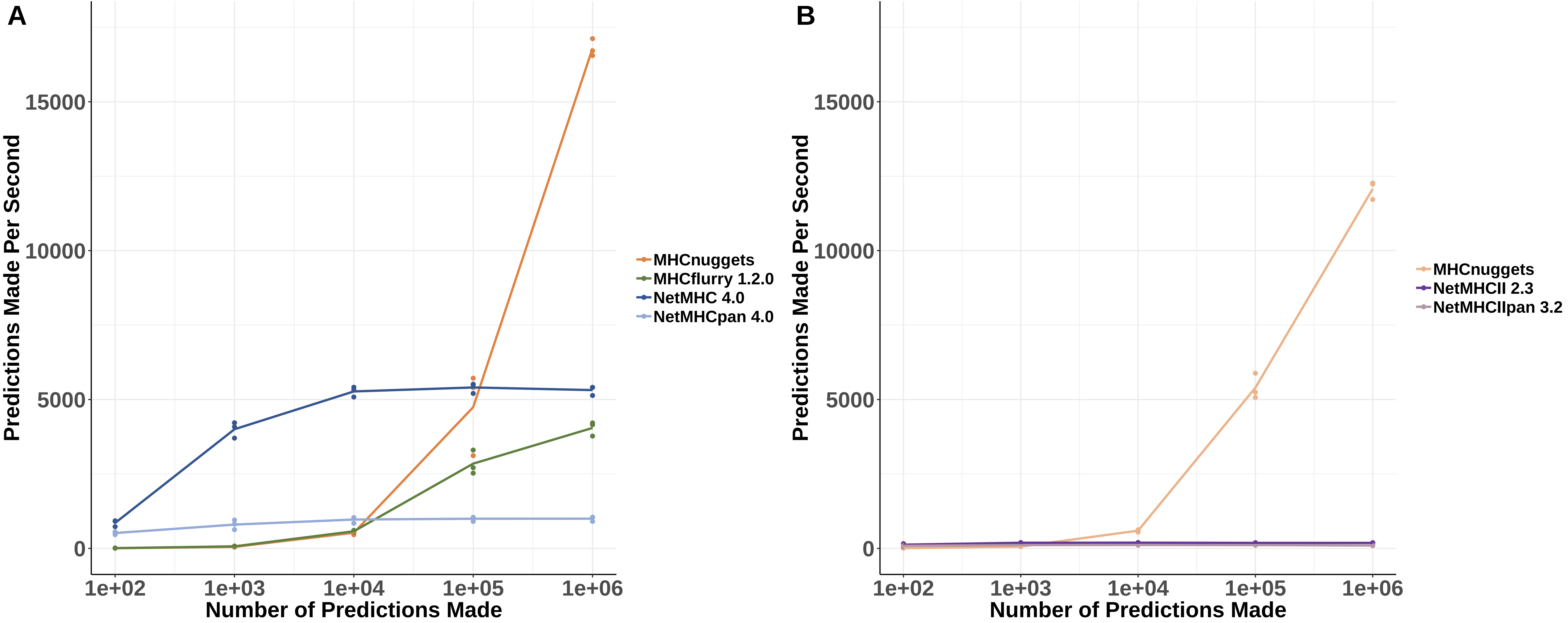

A
Input Sequence
A
K
N
…
D
N
X
LSTM Layer 64 units
LSTM Layer
LSTM Layer
LSTM Layer
…
LSTM Layer
LSTM Layer
LSTM Layer
FC Layer 64 units
…
Output Layer
Binding Prediction
C
Root Network
Transfer learning
Initial Network
Fine tuning
Final Network
B
8-mers
9-mers
10-mers
Peptide
AKNLYGMI
AKNLYGMIS
AKNLYGMISR
Input
AKNLYGMIXX……XX
X x 56
AKNLYGMISXX……XX
X x 55
AKNLYGMISRXX……XX
X x 54

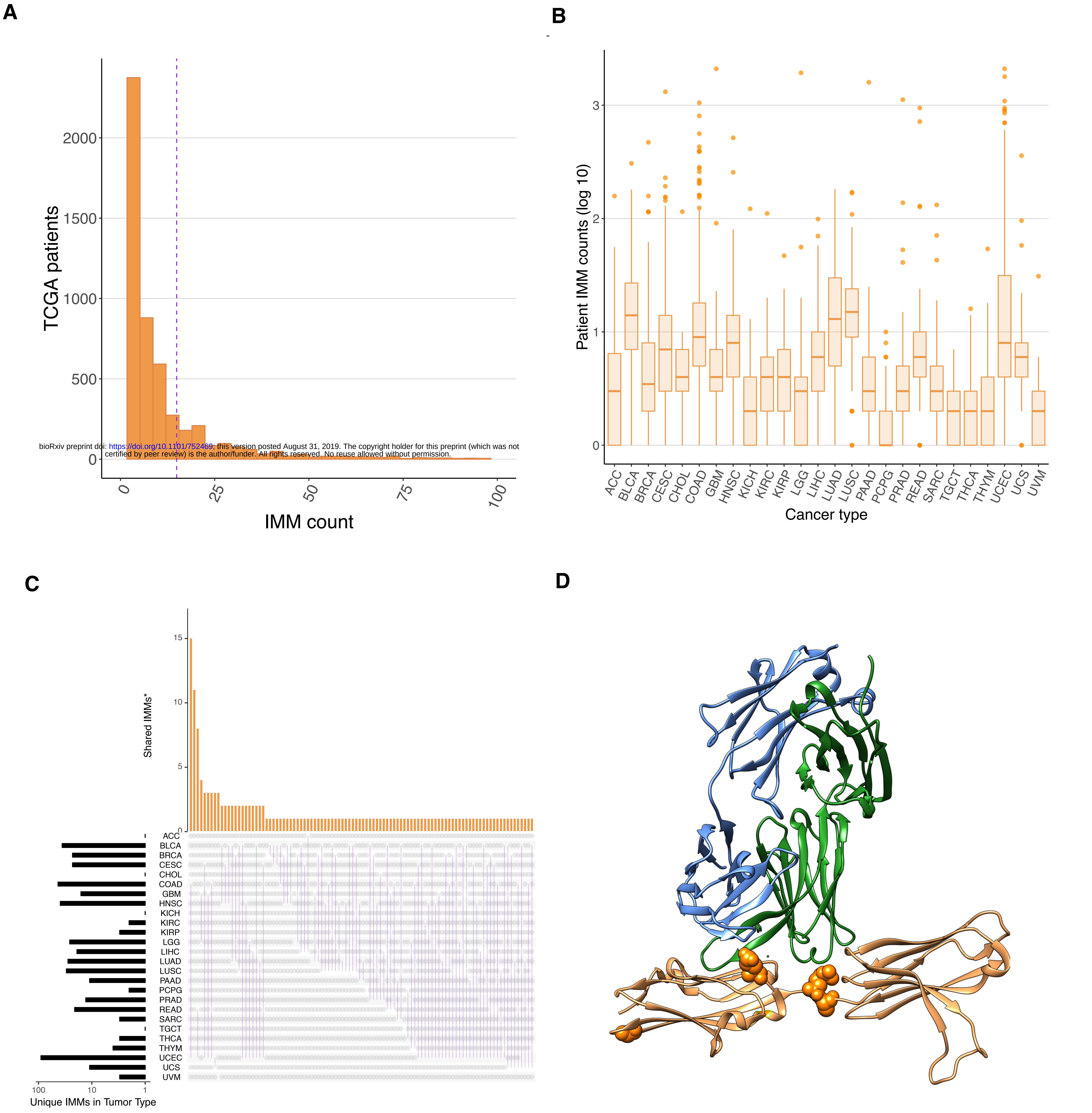

A
Class I Common Allele Prediction
MHCflurry
NetMHC
NetMHCpan
Class II Rare Allele Prediction
MHCnuggets
Class I Rare Allele Prediction
NetMHCIIpan
NetMHCII
Class II Common Allele Prediction
B
Binding Affinity Data
Mass Spectrometry Data
Training with Transfer Learning
Selection of Common or Rare Allele Model
Peptides of Interest
Binding Prediction


A
PPVn
1.00
0.75
0.50
0.25
0.00
Bonsack
Kim (BLIND)
MHCnuggets
MHCnuggets(no MS)
MHCflurry1.2.0
MHCflurry (no MS)
NetMHC4.0
NetMHCpan4.0
NetMHC3.0
NetMHCpan2.0
B
PPVn
1.00
0.75
0.50
0.25
0.00
A01:01
A02:01
A03:01
A24:02
A31:01
B51:01
Overall
MHCnuggets
MHCflurry1.2.0
MHCflurryMS
NetMHC4.0
NetMHCpan4.0
C
PPVn
1.00
0.75
0.50
0.25
0.00
8mers
9mers
10mers
11mers
12mers
MHCnuggets
MHCflurry1.2.0
MHCflurryMS
NetMHC4.0
NetMHCpan4.0
D
TP
250
200
150
100
50
0
FP
250
200
150
100
50
0
A01:01
A02:01
A03:01
A24:02
A31:01
B51:01
MHCnuggets
MHCflurry1.2.0
MHCflurryMS
NetMHC4
NetMHCpan4


A
PPVn
1.00
0.75
0.50
0.25
0.00
H−2−IAb
H−2−IAd
HLA−DPA1*01:03−DPB1*02:01
HLA−DPA1*01:03−DPB1*04:01
HLA−DPA1*02:01−DPB1*01:01
HLA−DPA1*02:01−DPB1*05:01
HLA−DPA1*03:01−DPB1*04:02
HLA−DQA1*01:01−DQB1*05:01
HLA−DQA1*01:02−DQB1*06:02
HLA−DQA1*03:01−DQB1*03:02
HLA−DQA1*04:01−DQB1*04:02
HLA−DQA1*05:01−DQB1*02:01
HLA−DQA1*05:01−DQB1*03:01
HLA−DRB1*01:01
HLA−DRB1*03:01
HLA−DRB1*04:01
HLA−DRB1*04:04
HLA−DRB1*04:05
HLA−DRB1*07:01
HLA−DRB1*08:02
HLA−DRB1*09:01
HLA−DRB1*11:01
HLA−DRB1*13:02
HLA−DRB1*15:01
HLA−DRB3*01:01
HLA−DRB4*01:01
HLA−DRB5*01:01
MHCnuggets
B
Score
1.00
0.75
0.50
0.25
0.00
auROC
KTAU
Pearson r
MHCnuggets
NetMHCII−2.3


A
Peptides of Interest for MHC Allele A
LOMO==FALSE
LOMO==TRUE
Model Network for MHC Allele A
Closest Model Network for MHC Allele A
MHCnuggets
Binding Prediction
B
PPVn
1.00
0.75
0.50
0.25
0.00
A03:19
A24:01
A32:07
A32:15
A68:23
B14:01
B15:16
B27:02
B27:04
B27:06
B27:09
B27:20
B39:06
B40:13
B44:01
B57:03
B58:02
B62:01
B81:01
C08:01
Average
MHCnuggets
C
PPVn
1.00
0.75
0.50
0.25
0.00
H–2–IAb
H–2–IAd
HLA–DPA10103–DPB10201
HLA–DPA10103–DPB10401
HLA–DPA10201–DPB10101
HLA–DPA10201–DPB10501
HLA–DPA10301–DPB10402
HLA–DQA10101–DQB10501
HLA–DQA10102–DQB10602
HLA–DQA10301–DQB10302
HLA–DQA10401–DQB10402
HLA–DQA10501–DQB10201
HLA–DQA10501–DQB10301
HLA–DRB10101
HLA–DRB10301
HLA–DRB10401
HLA–DRB10404
HLA–DRB10405
HLA–DRB10701
HLA–DRB10802
HLA–DRB10901
HLA–DRB11101
HLA–DRB11302
HLA–DRB11501
HLA–DRB30101
HLA–DRB40101
HLA–DRB50101
Average
MHCnuggets
D
auROC
1.00
0.75
0.50
0.25
0.00
H–2–IAb
H–2–IAd
HLA–DPA10103–DPB10201
HLA–DPA10103–DPB10401
HLA–DPA10201–DPB10101
HLA–DPA10201–DPB10501
HLA–DPA10301–DPB10402
HLA–DQA10101–DQB10501
HLA–DQA10102–DQB10602
HLA–DQA10301–DQB10302
HLA–DQA10401–DQB10402
HLA–DQA10501–DQB10201
HLA–DQA10501–DQB10301
HLA–DRB10101
HLA–DRB10301
HLA–DRB10401
HLA–DRB10404
HLA–DRB10405
HLA–DRB10701
HLA–DRB10802
HLA–DRB10901
HLA–DRB11101
HLA–DRB11302
HLA–DRB11501
HLA–DRB30101
HLA–DRB40101
HLA–DRB50101
Average
MHCnuggets
NetMHCIIpan3.2


A
Predictions Made Per Second
15000
10000
5000
0
1e+02
1e+03
1e+04
1e+05
1e+06
Number of Predictions Made
MHCnuggets
MHCflurry 1.2.0
NetMHC 4.0
NetMHCpan 4.0
B
Predictions Made Per Second
15000
10000
5000
0
1e+02
1e+03
1e+04
1e+05
1e+06
Number of Predictions Made
MHCnuggets
NetMHCII 2.3
NetMHCIIpan 3.2


A
TCGA patients
2000
1500
1000
500
0
IMM count
0
25
50
75
100
B
Patient IMM counts (log 10)
3
2
1
0
Cancer type
ACC
BLCA
BRCA
CESC
CHOL
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
PAAD
PCPG
PRAD
READ
SARC
TGCT
THCA
THYM
UCEC
UCS
UVM
C
Shared IMMs*
15
10
5
0
ACC
BLCA
BRCA
CESC
CHOL
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
PAAD
PCPG
PRAD
READ
SARC
TGCT
THCA
THYM
UCEC
UCS
UVM
100
10
1
Unique IMMs in Tumor Type
D