Neopepsee:

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

Chris Milianta, Stefania del Rosario, Jovonny Trinh

What are neoantigens?

- Neopeptide fragments that are induced by somatic mutations, some of which can induce a T-cell response
- Only found in cancer/tumor cells, which is why they have recently become targets for immunotherapy
- They are presented on the surface of a tumor cell by MHC-I molecules

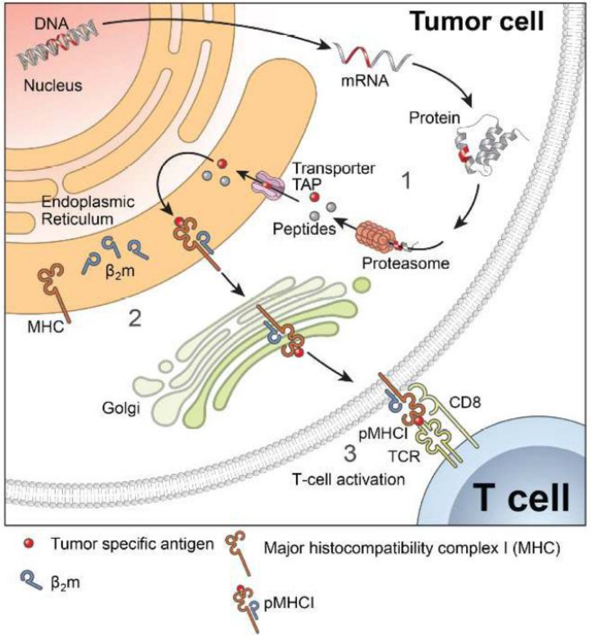


Figure 1A

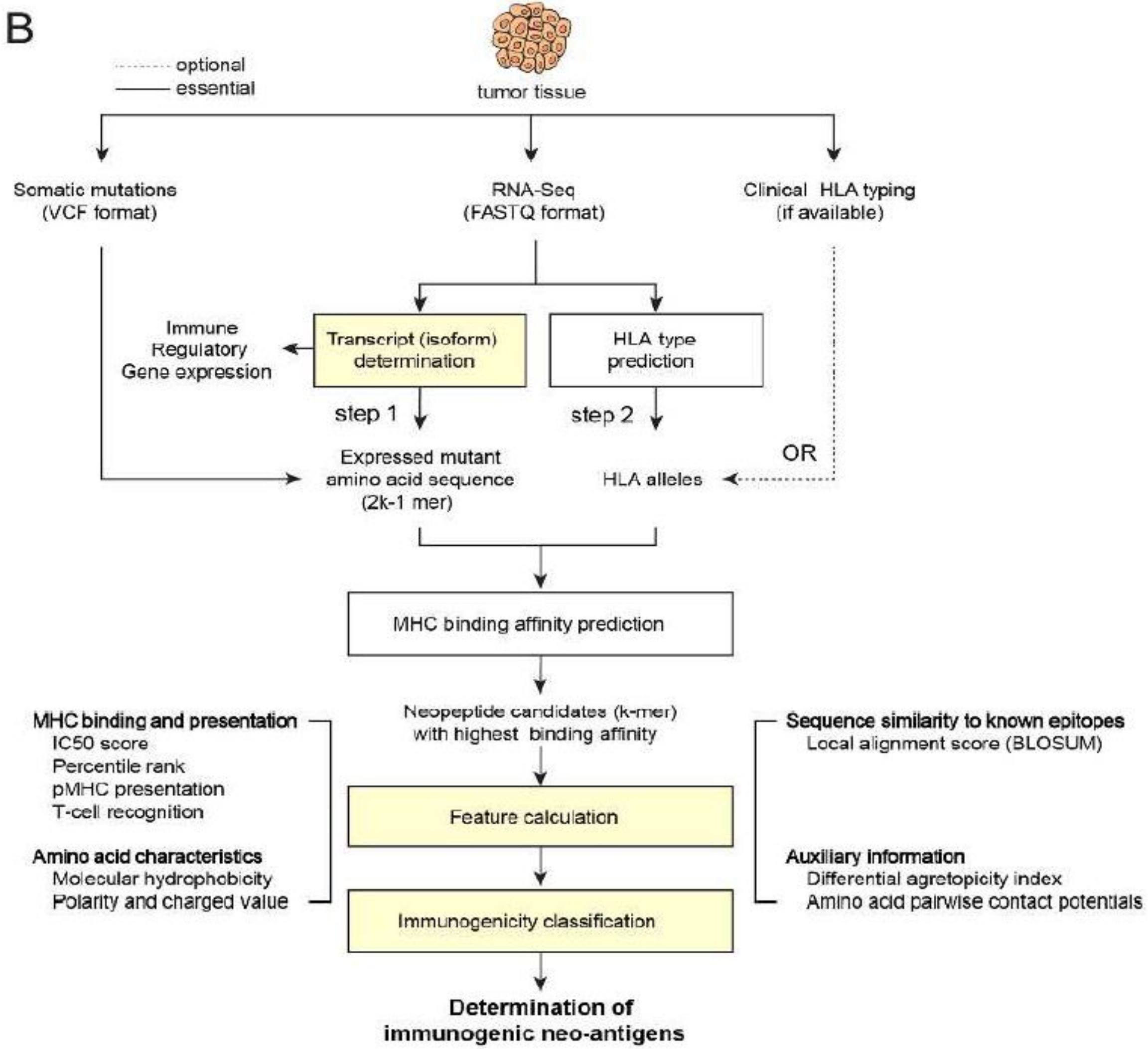
Current Issues with Neoantigen Prediction

- Large number of false-positive predictions of Neoantigens exist
- Genome level applications are limited
  - Relies on arbitrary cut-offs of predicted MHC-I binding affinities.
  - Some genes are not included in analysis. Ex. isoform-specific gene epression, immune signature-related, etc...
  - Analysis requires complex computation processes reserved for bioinformatics experts

What is Neopepsee?

- A machine-learning based program that was constructed from 14 immunogenicity features
- Aim is to enhance accuracy in predicting neoantigens
- Automatically selects mutated peptide sequences from raw RNA-sequence data and a compiled list of somatic mutations
- Was tested on melanoma, leukemia, and stomach cancer data sets

Overall Workflow of Neopepsee



B

optional

essential

tumor tissue

Somatic mutations (VCF format)

RNA-Seq (FASTQ format)

Clinical HLA typing (if available)

Immune Regulatory Gene expression

Transcript (isoform) determination

HLA type prediction

step 1

step 2

OR

Expressed mutant amino acid sequence (2k-1 mer)

HLA alleles

MHC binding affinity prediction

Neopeptide candidates (k-mer) with highest binding affinity

MHC binding and presentation
- IC50 score
- Percentile rank
- pMHC presentation
- T-cell recognition

Amino acid characteristics
- Molecular hydrophobicity
- Polarity and charged value

Sequence similarity to known epitopes
- Local alignment score (BLOSUM)

Auxiliary information
- Differential agretopicity index
- Amino acid pairwise contact potentials

Feature calculation

Immunogenicity classification

Determination of immunogenic neo-antigens

Methods: Neopepsee parameters

Identification of potential predictors for immunogenicity:

14 predictors in 3 categories

(A) MHC-binding and presentation

(B) Amino acid characteristics

(C) Complex scores

Table S1. Potential features for immunogenicity prediction.

Feature	Description
A. MHC binding and presentation	
_*IC50_	Binding affinity between peptide and MHC
*Rank	Percentile rank generated by comparing the peptide's _IC50_ against those of a set of random peptides from SWISSPROT database.
MHC score	MHC class I binding affinity into a MHC class I pathway likelihood score
TAP score	Prediction of transporter associated with antigen processing transport efficiency
Cleavage score	Prediction of proteasomal cleavage
*Combined score	Combined prediction score
*Immunogenicity score	Relative ability of a peptide/MHC complex to elicit an immune response
B. Amino acid characteristics	
*Hydrophobicity	Hydrophobicity score of the peptide
*Polarity & Charged score	Polarity and charged score of the peptide
Molecular size	Molecular size of the peptide
Entropy	The sum of entropy of each amino acid
C. Sequence similarity to known epitopes and other auxiliary information	
#DAI	_IC50 difference between wild-type and mutant peptides_
*AAPPs	Connectivity between the peptide and binding site of the MHC
*Similarity score	Similarity score between the peptide and known epitopes

Methods: Neopepsee Parameters

Positive set:

Epitopes that exhibited positive T-cell response in humans & were highly conserved to human proteins in Swiss-Prot

Negative set:

Randomly selected Peptides from Single Nucleotide Polymorphisms to conserve naturally occurring ratio of neoantigens (1:48)

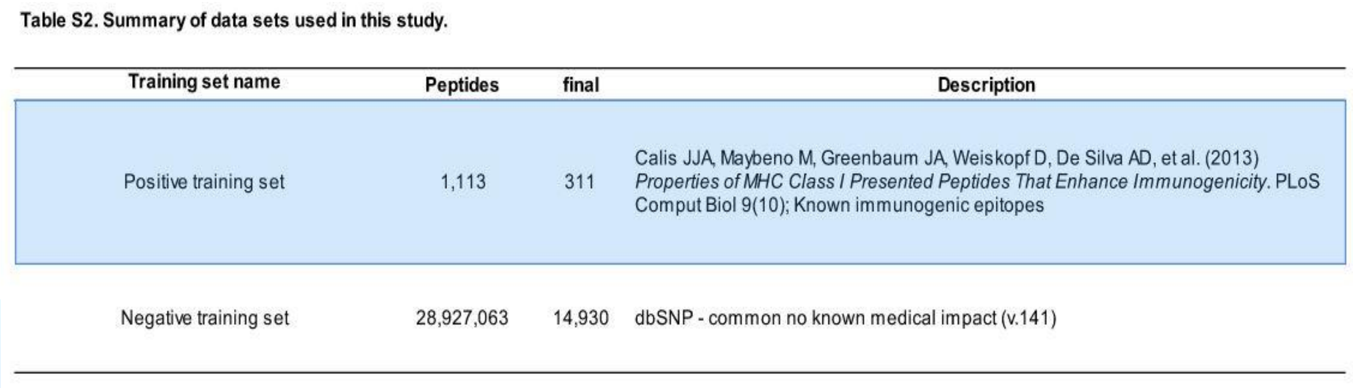
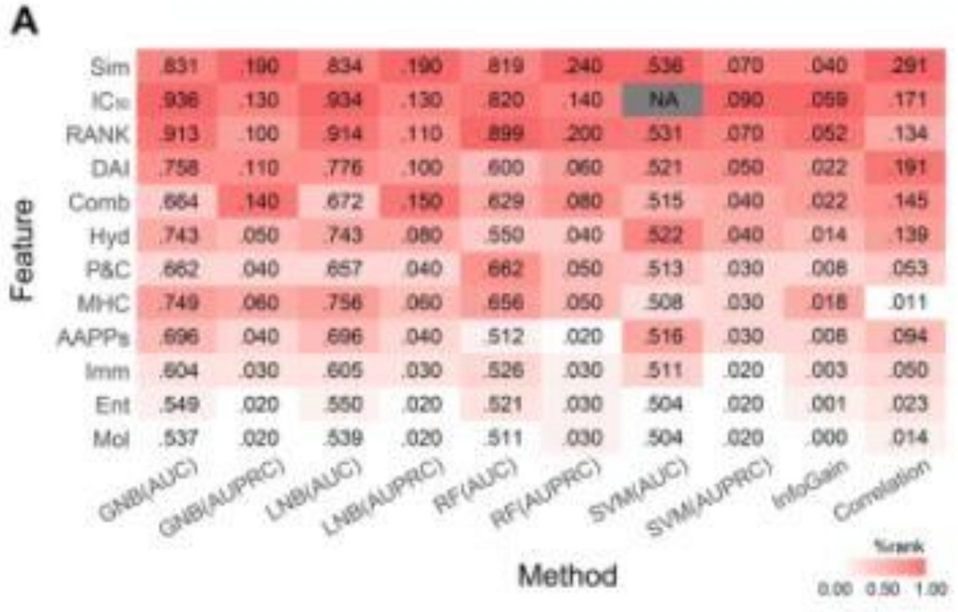


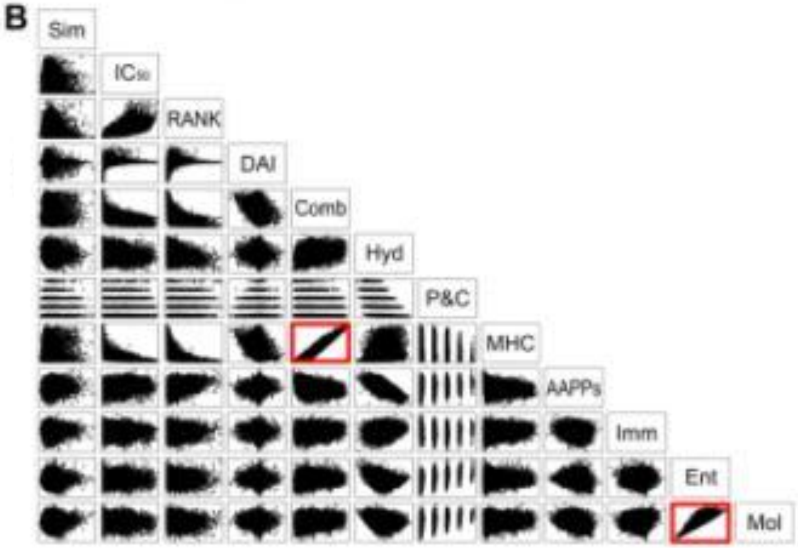
Table S2. Summary of data sets used in this study.

Training set name	Peptides	final	Description
Positive training set	1,113	311	Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, et al. (2013) *Properties of MHC Class I Presented Peptides That Enhance Immunogenicity*. PLoS Comput Biol 9(10); Known immunogenic epitopes
Negative training set	28,927,063	14,930	dbSNP - common no known medical impact (v.141)

Methods: Feature Selection

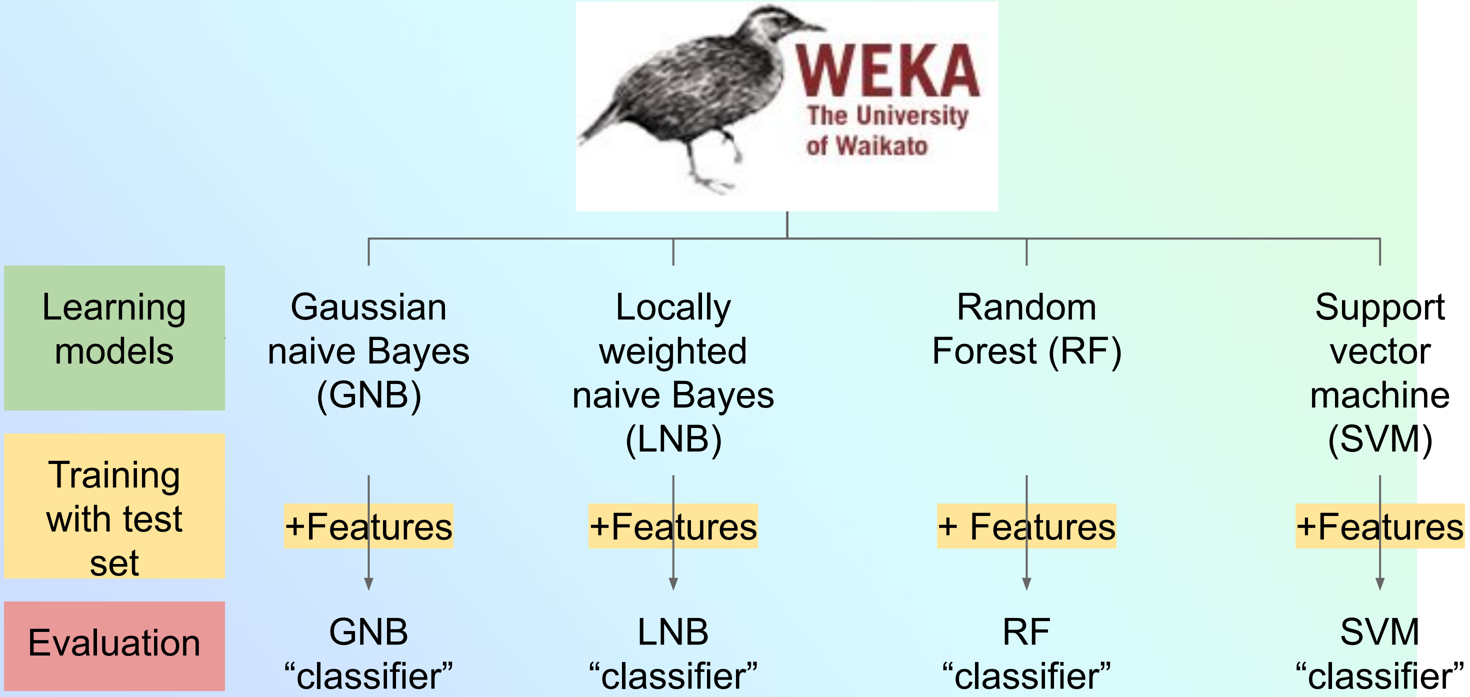


	GNB(AUC)	GNB(AUPRC)	LNB(AUC)	LNB(AUPRC)	RF(AUC)	RF(AUPRC)	SVM(AUC)	SVM(AUPRC)	InfoGain	Correlation
Sim	.831	.190	.834	.190	.819	.240	.536	.070	.040	.291
IC_{50}	.936	.130	.934	.130	.820	.140	NA	.090	.059	.171
RANK	.913	.100	.914	.110	.899	.200	.531	.070	.052	.134
DAI	.758	.110	.776	.100	.600	.060	.521	.050	.022	.191
Comb	.664	.140	.672	.150	.629	.080	.515	.040	.022	.145
Hyd	.743	.050	.743	.080	.550	.040	.522	.040	.014	.139
P&C	.662	.040	.657	.040	.662	.050	.513	.030	.008	.053
MHC	.749	.060	.756	.060	.656	.050	.508	.030	.018	.011
AAPPs	.696	.040	.696	.040	.512	.020	.516	.030	.008	.094
Imm	.604	.030	.605	.030	.526	.030	.511	.020	.003	.050
Ent	.549	.020	.550	.020	.521	.030	.504	.020	.001	.023
Mol	.537	.020	.539	.020	.511	.030	.504	.020	.000	.014



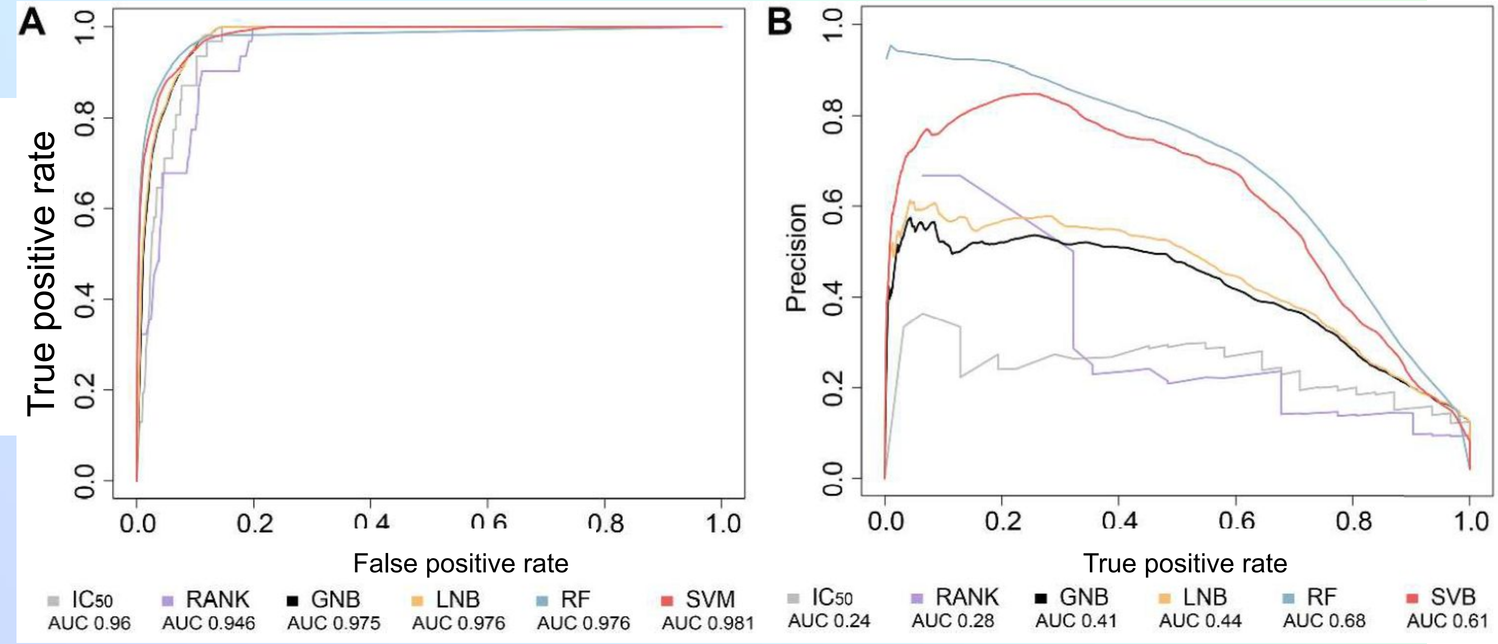
- IC50
- Rank
- Combined score
- Immunogenicity score
- Hydrophobicity
- Polarity and charged score
- DAI
- AAPPs
- Similarity

Methods: Machine learning based classification



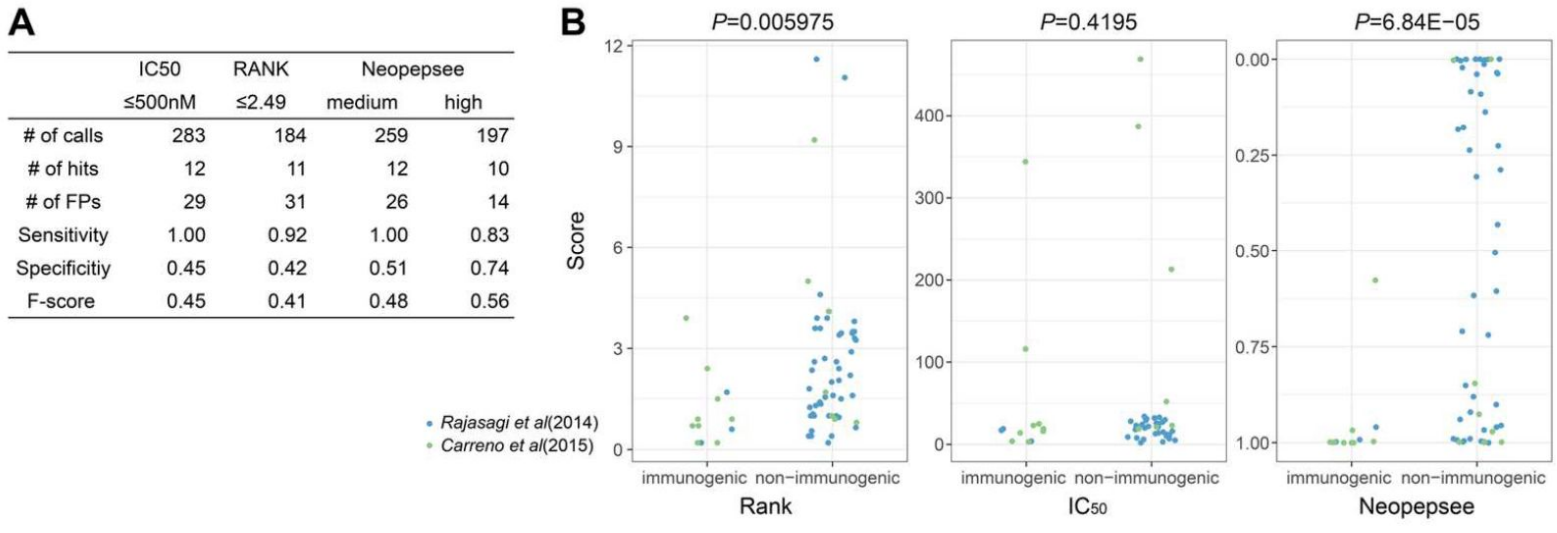
Learning models	Gaussian naive Bayes (GNB)	Locally weighted naive Bayes (LNB)	Random Forest (RF)	Support vector machine (SVM)
Training with test set	+Features	+Features	+ Features	+Features
Evaluation	GNB "classifier"	LNB "classifier"	RF "classifier"	SVM "classifier"

Results: Evaluation of Neoantigen Prediction in Neopepsee



Area under curve = probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

Results: Tests on Independent Experimental Data

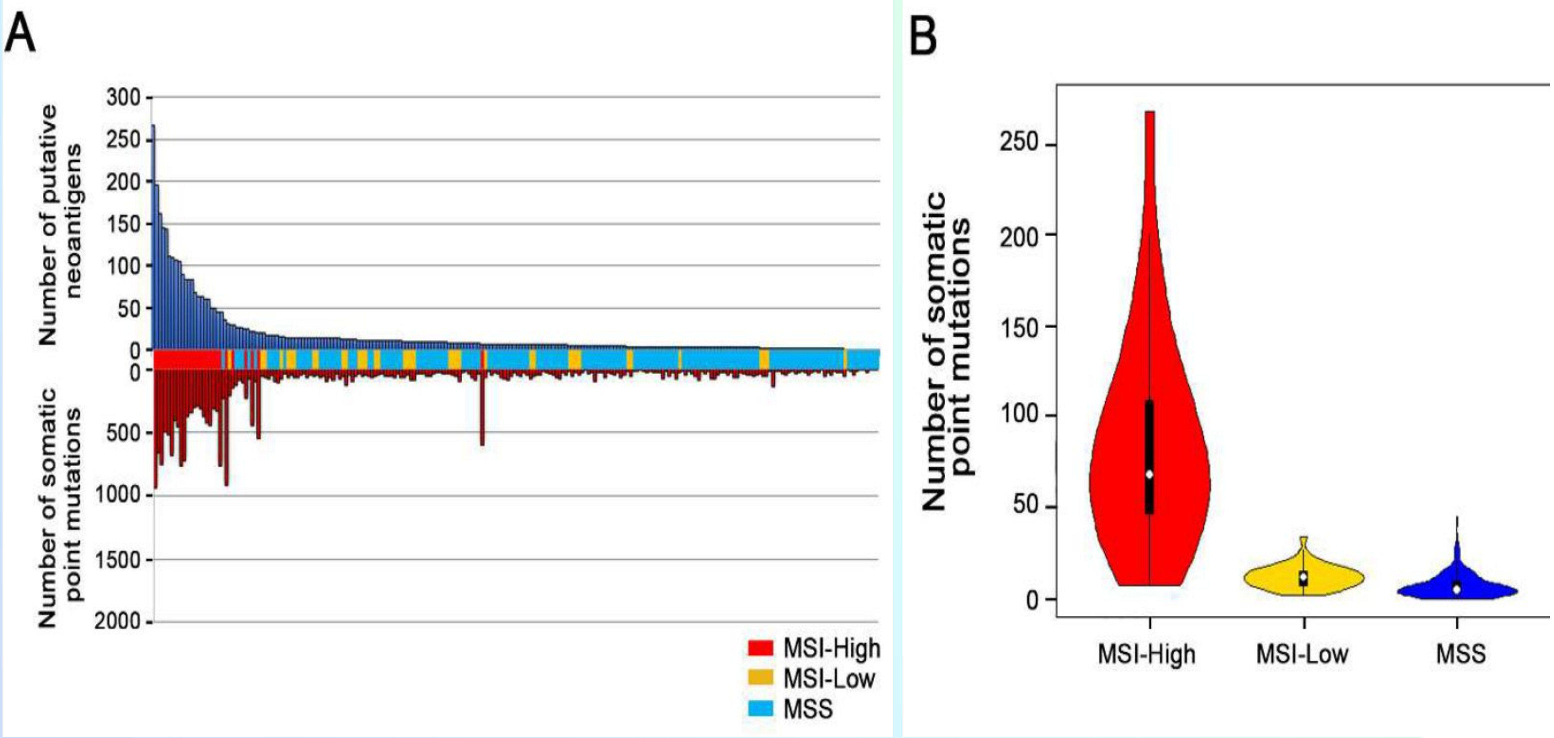


A

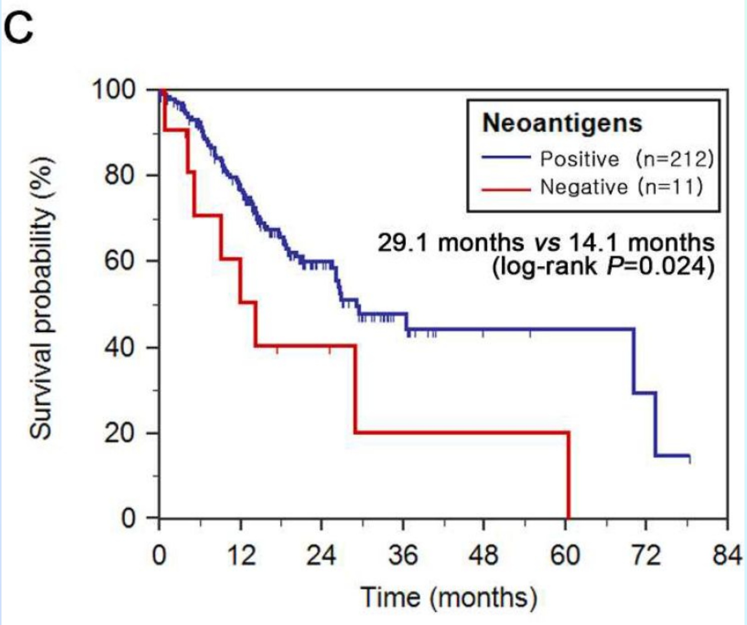
	IC50	RANK	Neopepsee	
	≤500nM	≤2.49	medium	high
# of calls	283	184	259	197
# of hits	12	11	12	10
# of FPs	29	31	26	14
Sensitivity	1.00	0.92	1.00	0.83
Specificitiy	0.45	0.42	0.51	0.74
F-score	0.45	0.41	0.48	0.56

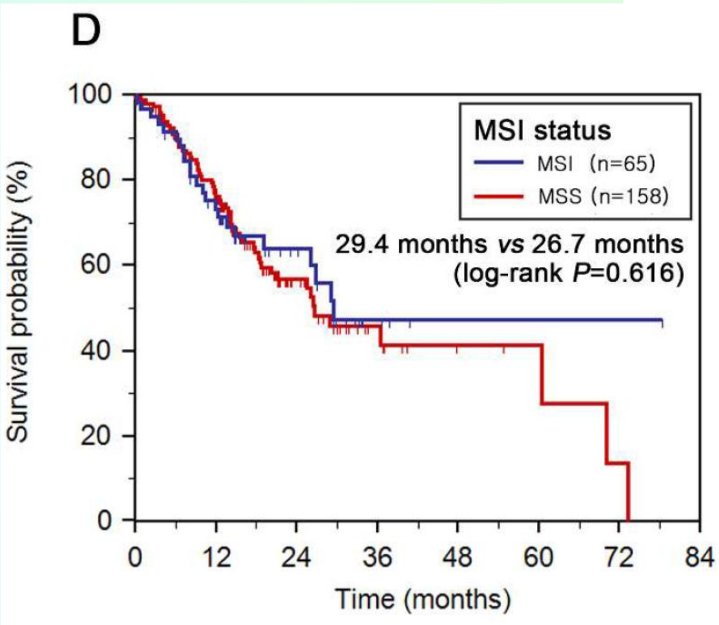
Neopepsee compared to traditional methods for neoantigen predictions.

Results: Application to the TCGA-STAD Dataset

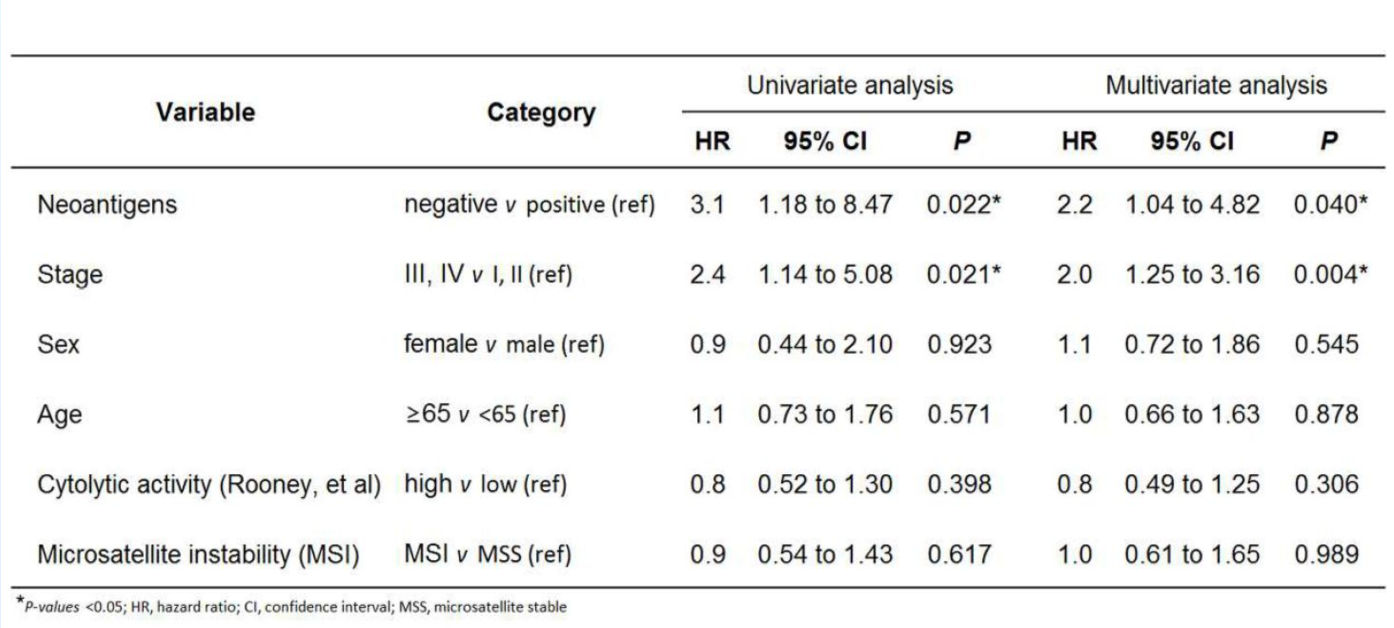


Results: Application to the TCGA-STAD Dataset...





Results: Application to the TCGA-STAD Dataset...



Variable	Category	Univariate analysis			Multivariate analysis		
		HR	95% CI	*P*	HR	95% CI	*P*
Neoantigens	negative *v* positive (ref)	3.1	1.18 to 8.47	0.022*	2.2	1.04 to 4.82	0.040*
Stage	III, IV *v* I, II (ref)	2.4	1.14 to 5.08	0.021*	2.0	1.25 to 3.16	0.004*
Sex	female *v* male (ref)	0.9	0.44 to 2.10	0.923	1.1	0.72 to 1.86	0.545
Age	≥65 *v* <65 (ref)	1.1	0.73 to 1.76	0.571	1.0	0.66 to 1.63	0.878
Cytolytic activity (Rooney, et al)	high *v* low (ref)	0.8	0.52 to 1.30	0.398	0.8	0.49 to 1.25	0.306
Microsatellite instability (MSI)	MSI *v* MSS (ref)	0.9	0.54 to 1.43	0.617	1.0	0.61 to 1.65	0.989

*P-values <0.05; HR, hazard ratio; CI, confidence interval; MSS, microsatellite stable

Conclusion

- Efficient method to maximize neoantigen predictions.
- Neopepsee can be applied to identify putative neoantigens, and can also be used to compare neoantigens with known immune epitopes. The analysis results can be used for prognostic/predictive biomarker discovery or to design antigens for cancer vaccines.

Relevance

- Goals for neoantigen prediction software is to classify patients who will benefit from immunotherapy, or to design a personalized cancer vaccine.
- Neopepsee will enable the efficient analysis personal somatic mutation profiles and identify potential neopeptides for personalized vaccination.