Predicting peptides presentation by major histocompatibility class I: an improved machine learning approach to the immunopeptidome

•••

Presentation by: Hsuan-Yeh Pan and Josua Aponte-Serrano

Outline

1. Biological Background
2. Data Collection
3. Machine Learning Approaches
4. Review of Random Forest
5. Objectives, Data and Features
6. Metrics
7. Results
    a. Feature Importance
    b. Performance Plot
    c. Information Content
    d. New Data
    e. Chemical Affinity and Gene Expression Data

Biological Background (I)

- Objective: to characterize and predict peptides presented by Major Histocompatibility Complex I
- Genes that code for proteins found on the surfaces of cells that help the immune system recognize foreign substances
- Important components of the immune system because they allow T lymphocytes to recognize defective cells

There are two types of MHC molecule, MHC class I and MHC class II.

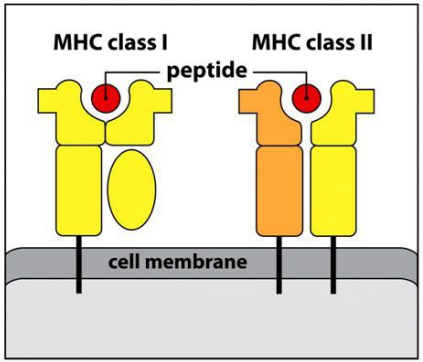
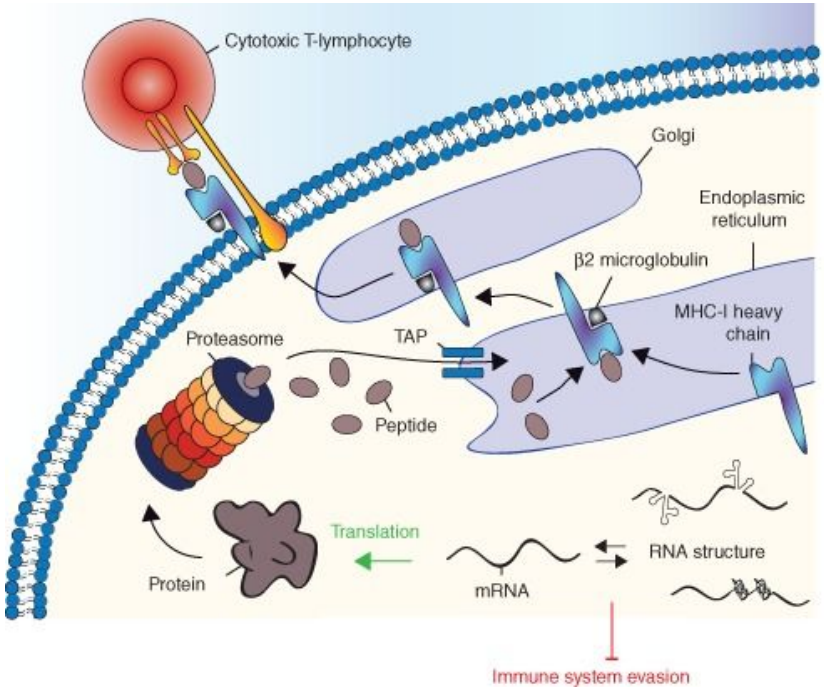


Figure 3.8 The Immune System, 3ed. (© Garland Science 2009)

Biological Background (II)

- Targeted proteins are degraded by the proteosome, transported to the ER where they bind to MHCs and exported to the membrane
- Types of peptides presented by MHCs
  - House-keeping proteins
  - Viral proteins (infected)
  - Neoplastic cell proteins



Immunopeptidome Data Collection

- MHCs can be purified using different techniques. In this paper, cells are lysed, MCHs are captured by a monoclonal antibody and eluted using affinity chromatography
- Chemical affinity data can be collected using different biochemical assays (example: quantitative ELISA)
  - Requires peptide synthesis and selection
  - Peptide presentation is not determined solely by chemical affinity
- Mass Spectrometry (MS) characterizes chemical compounds in a sample by sorting ions according to their mass
  - Describes the peptides presented *in vivo*
  - Captures information other than chemical affinity: half-life, proteasomal processing and abundance of protein sequences

Machine Learning Approaches

- Using MS to characterize the immunopeptidome for clinical applications is costly and requires large samples from patients
- Machine Learning approaches have been used to predict peptide presentation
  - Based on Artificial Neural Networks:
    - NetMHC - trained in chemical affinity data
    - NetMHCstabpan - trained on data of the half-life of the MHC-peptide complex in vitro
    - NetMCHpan - trained in chemical affinity and MS data
  - Based on Position Weight Matrices:
    - MixMHCpred - trained in MS data
- This paper presents a new random forest classifier (ForestMCH) and compares performance with previous ML approaches

Random Forest (I)

- Randomly select data from original dataset to make bootstrapped dataset
  - The bootstrapped dataset is same size as original dataset
  - The bootstrapped dataset can pick same data from original dataset

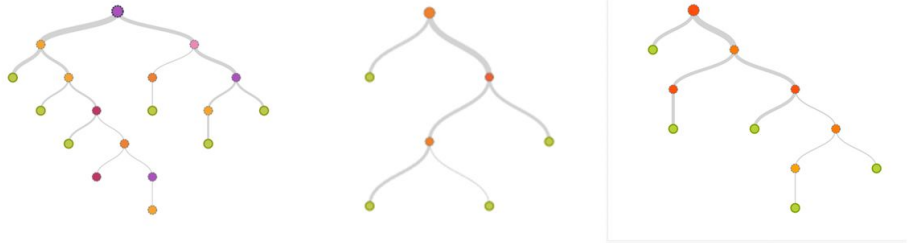
Hydropathy (hydrophobic or hydrophilic properties)	Presence of aromatic	Charge at physiological pH	Mass	....etc.	label
hydrophobic	Yes	Neutral	400		A01
hydrophobic	No	Positive	397		A02
hydrophilic	Yes	Positive	200		B02
hydrophobic	Yes	Negative	333		B03
hydrophilic	no	positive	345		A03

Hydropathy (hydrophobic or hydrophilic properties)	Presence of aromatic	Charge at physiological pH	Mass	....etc.	label
hydrophobic	Yes	Neutral	400		A01
hydrophobic	No	Positive	397		A02
hydrophilic	Yes	Positive	200		B02
hydrophilic	no	positive	345		A03
hydrophilic	no	positive	345		A03

Random Forest (II)

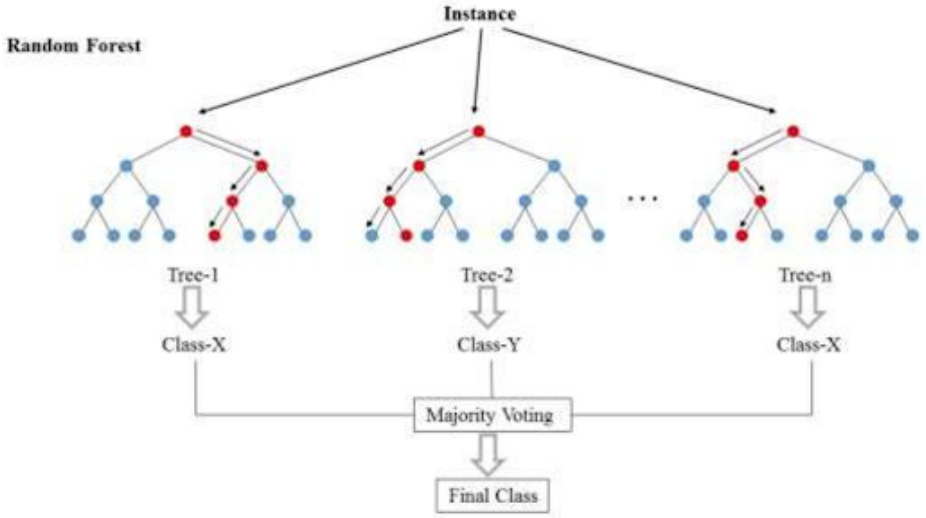
- Create decision tree using the bootstrapped dataset
  - Only use a random subset of variables (features) for each node
- Create multiple trees by repeating previous steps (1000 trees in the paper)

Hydropathy (hydrophobic or hydrophilic properties)	Presence of aromatic	Charge at physiological pH	Mass	....etc.	label
hydrophobic	Yes	Neutral	400		A01
hydrophobic	No	Positive	397		A02
hydrophilic	Yes	Positive	200		B02
hydrophilic	no	positive	345		A03
hydrophilic	no	positive	345		A03



Random Forest (III)

- Using remain data from original dataset to test if the random forest accuracy
  - Normally 1/3 of data in original dataset is not use for creating bootstrapped dataset and decision trees (a.k.a. out-of-bag data)
- Change number of variables and make decision trees again
  - compare how many variables for nodes can get most accuracy



Objective, Data and Features

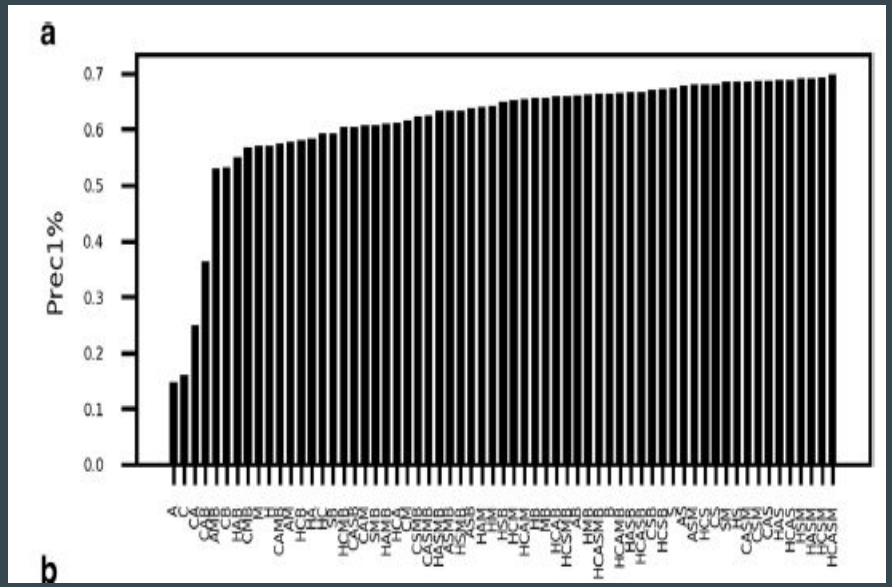
- Objective: to compare different ML approaches for characterizing and predicting peptide presentation
- Methods:
  - Immunopeptide data from 24 different datasets
  - Polyallelic samples deconvoluted using MixMHCpred
  - 1.6E5 nonamers assigned to 82 alleles
  - Training set: a 1:1 ratio of randomly generated nonamers from SwissProt to true binders
  - Test set: 99:1 ratio of random decoys to true binders (unbalanced data)
- Features:
  - Hydropathy, Blosum62 sequence encoding, One-hot (sparse) sequence encoding, Aromaticity, Mass, and Charge at physiological pH

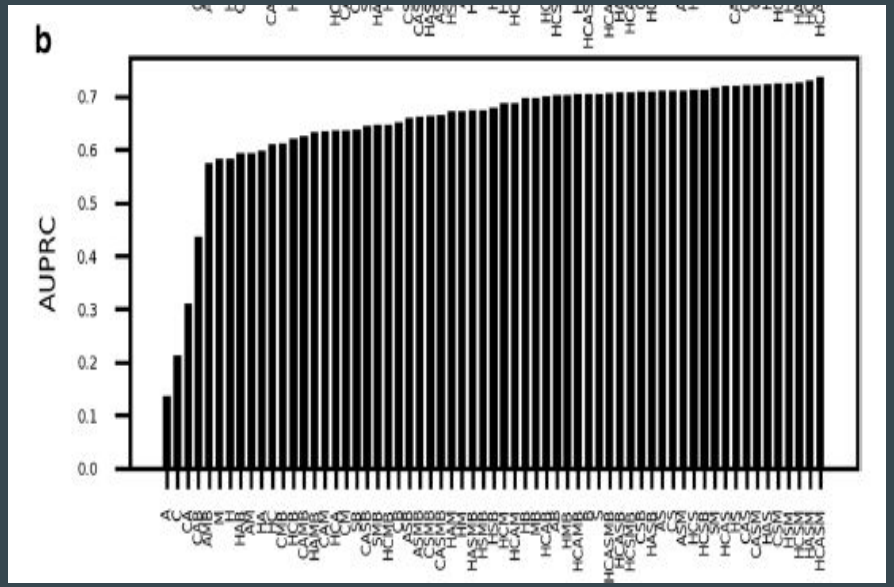
Metrics

- Classifiers return a ranking of peptides by MHC presentation
- Precision at 1%
  - Top 1% predicted positives, remaining 99% predicted negatives
  - Measures how many true positives among predicted positives
  - Values: 1.0 for a perfect classifier and 0.01 for a random classifier
- Area under the Precision Recall Curve (AUPRC)
  - AUPRC: true positives among predicted positives for different cutoffs values
  - Values: 1.0 for a perfect classifier and 0.01 for a random classifier
- Gini impurity
  - Measures the probability of an element to be incorrectly label

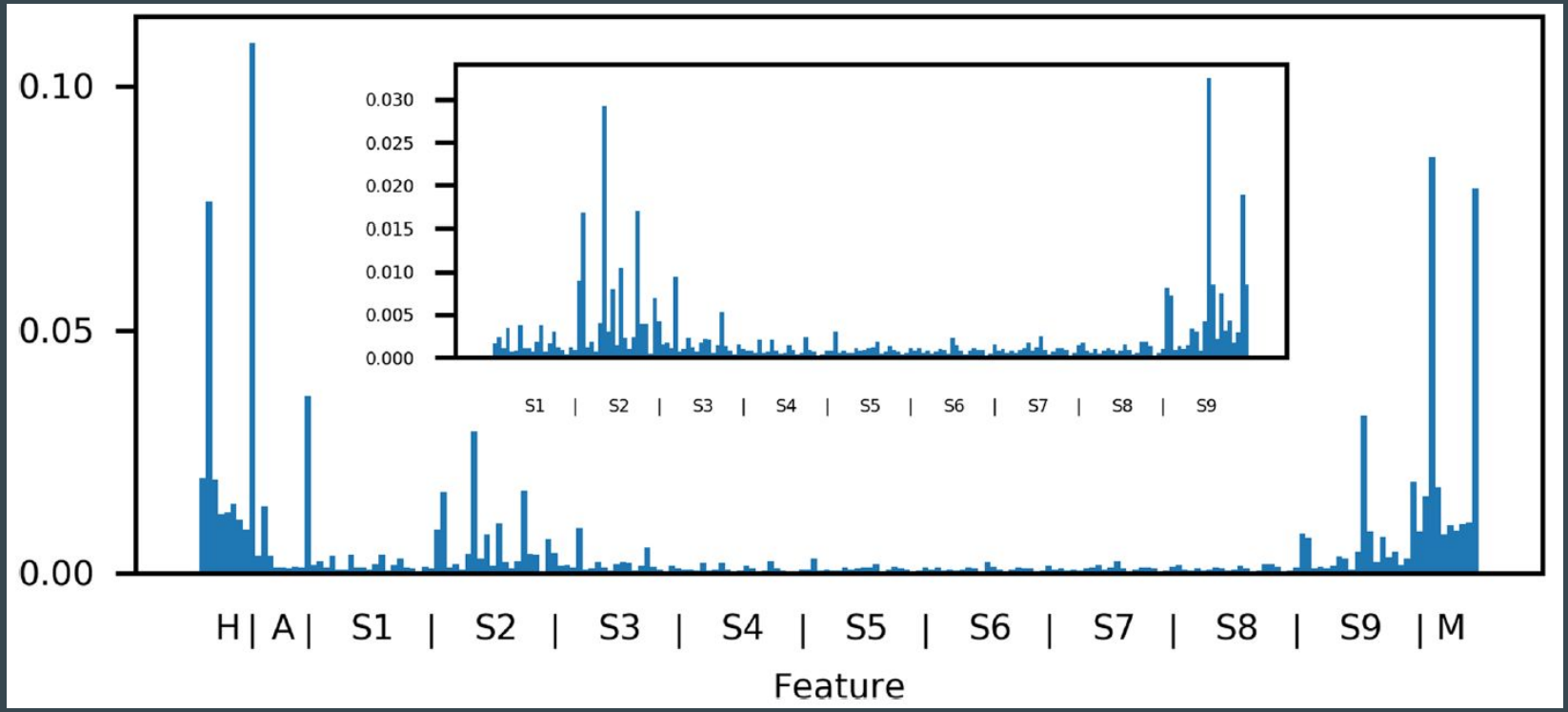
PAPER RESULTS

Results 1: Feature Selection by Performance

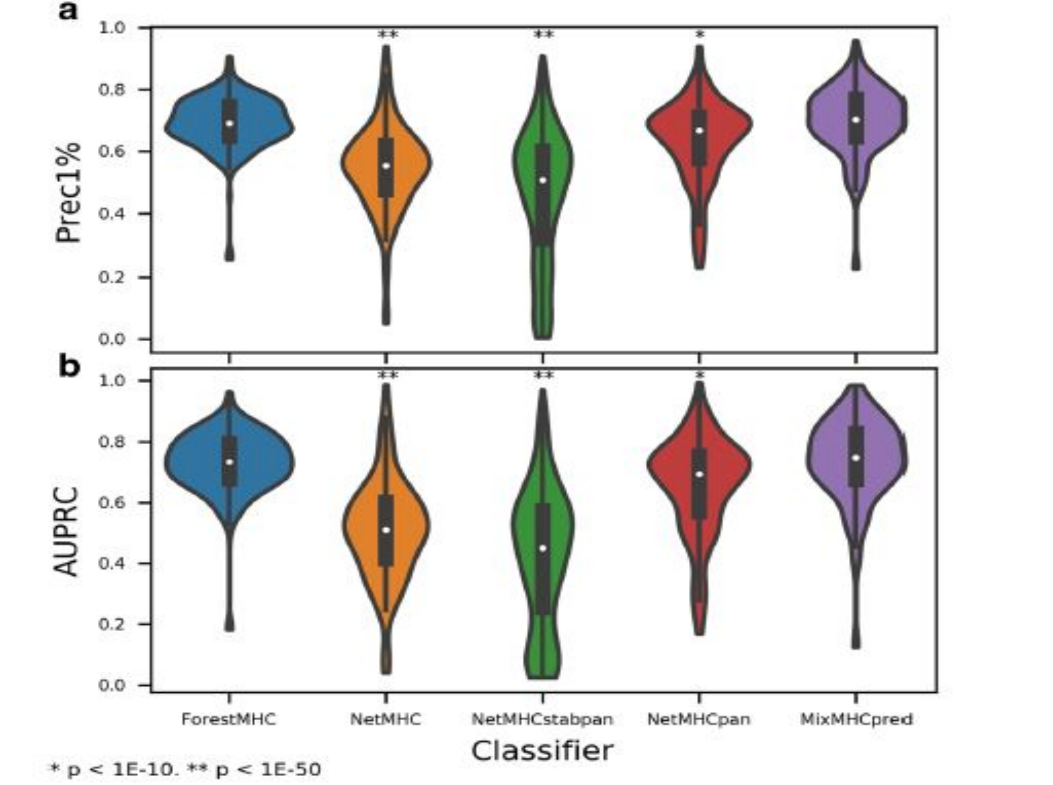




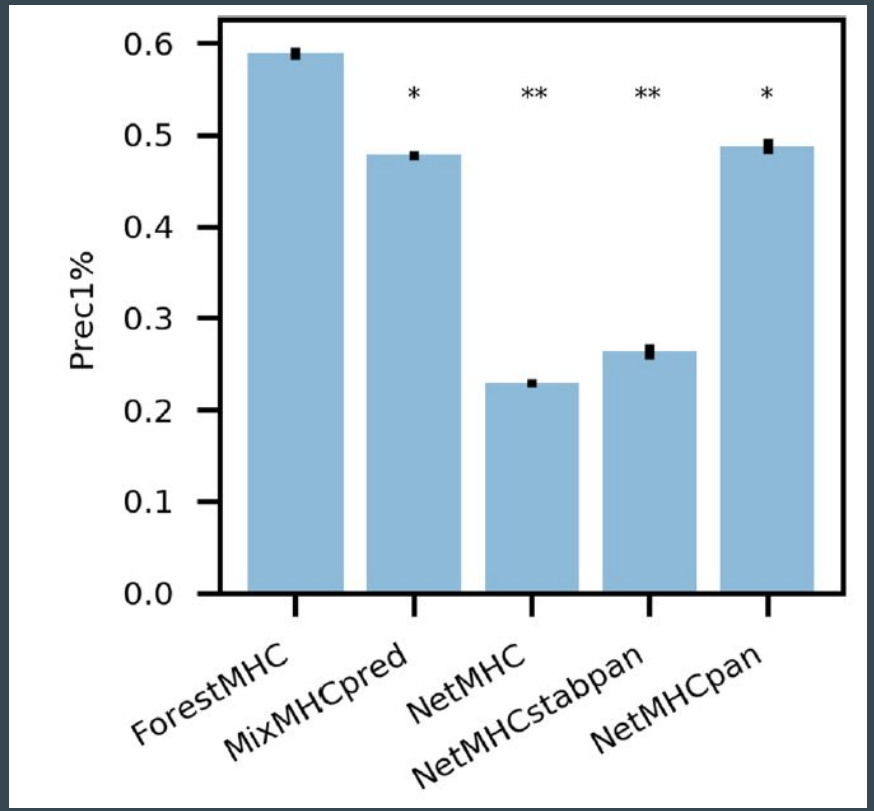
Results 2: Information Content by Feature



Results 3: Comparison of Classifiers using Test Data

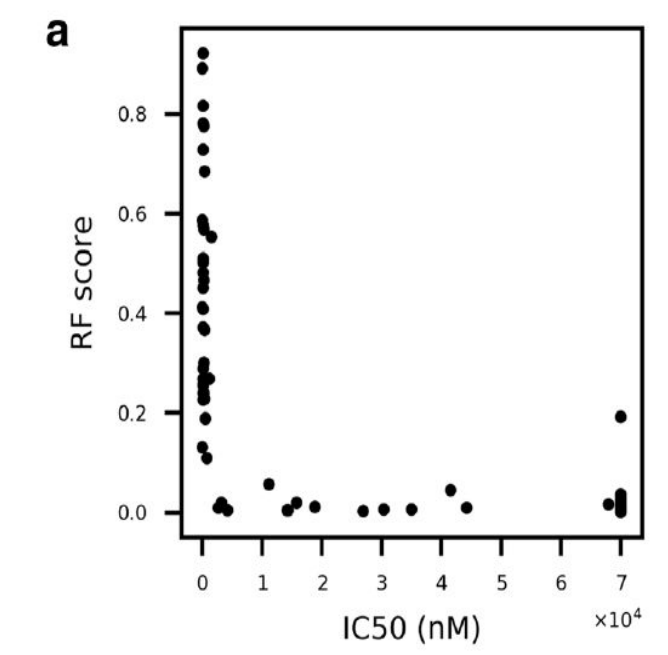


Results 4: Validation on Never-Before-Seen Data

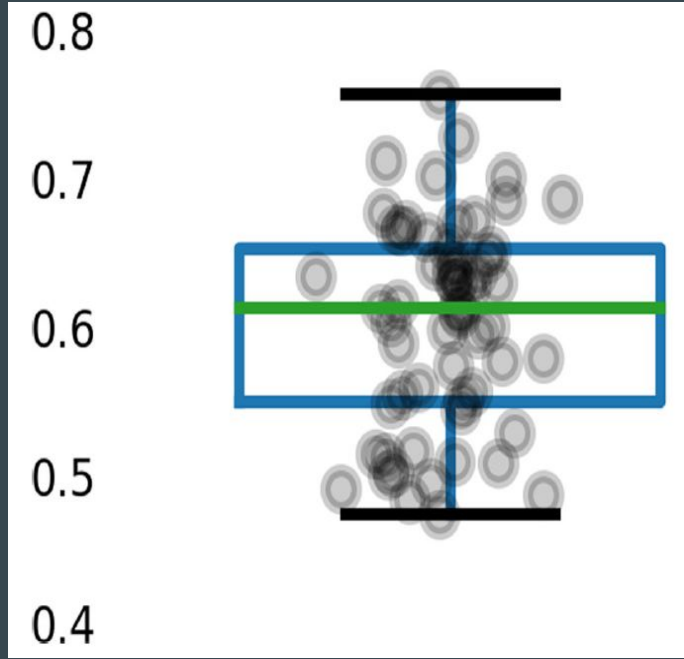


Results 5

Correlation RF score and chemical affinity data



Correlation of Gene Expression and MHC presentation



Conclusions

- ForestMHC yields greater precision than NetMHC and NetMHCpan and performs indistinguishably from MixMHCpred
  - MixMHCpred was used for deconvolution of polyallelic datasets
  - MixMHCpred and ForestMHC trained on same data
- ForestMHC outperforms MixMHCpred, NetMHC and NetMHCpan when tested on new ovarian carcinoma data
- Lack of linear correlation between MS and chemical affinity data
  - In vivo presentation only partially dependent on chemical affinity
  - Other explicative factors within MS data: positive effect of gene expression on presentation
- Identifying peptides presented by MHC-I is critical to extend our knowledge of the immunopeptidome and for clinical applications such as neoantigen-based cancer immunotherapy