BERTMHC: Improved MHC-peptide class II interaction prediction with transformer and multiple instance learning

Jun Cheng, Kaïdre Bendjama, Karola Rittner, Brandon Malone

1 Supplementary Methods

1.1 Data creation for patient mass spectrometry

We collected tumoral samples from **six** patients diagnosed with Non-Small Cell Lung Cancer (NSCLC) who were eligible for surgical resection. Tumor samples were rapidly snap frozen on liquid nitrogen upon collection. The tissue samples were lysed according to standard procedure with an Ultra Turrax (IKA Werke). The amount of tumor tissue was between 105 and 365 mg. From each of the tumor tissue lysates, HLA DR complexes were purified and analyzed by high resolution data-dependent acquisition mass spectrometry on a Q Exactive HF-X.

In order to identify peptides bound to the HLA DR molecules, resulting data was searched with MaxQuant (version 1.6.3.4) against a database of human protein sequences from the reference proteome downloaded 26th June 2020 from UniProt (UniProt Consortium, 2019) website and containing 75 004 protein sequences. MaxQuant was run with the following setting: digestion mode: unspecific; first search 20 ppm then search 4.5 ppm; fragment mass tolerance: 20 ppm; one variable modification (oxidation of methionine); no specific amino acids for the creation of decoy databases; peptide FDR 5%; protein FDR 100%; peptide length allowed: from 8 to 25 amino acids; reverse hits and contaminants were eliminated.

The Class-II HLA types of each patient was determined by `seq2hla` tool (Boegel *et al.*, 2013) from RNA-seq data from tumor samples.

References

Boegel, S., Löwer, M., Schäfer, M., Bukur, T., De Graaf, J., Boisguérin, V., Türeci, Ö., Diken, M., Castle, J. C., and Sahin, U. (2013). HLA typing from RNA-Seq sequence reads. *Genome Medicine*, **4**(12), 102.

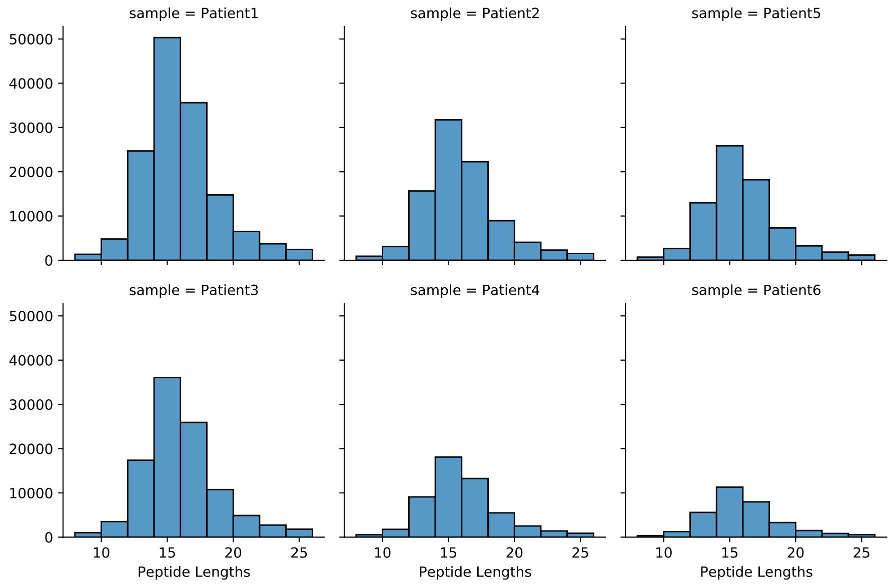
UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515.

2 Supplementary Tables

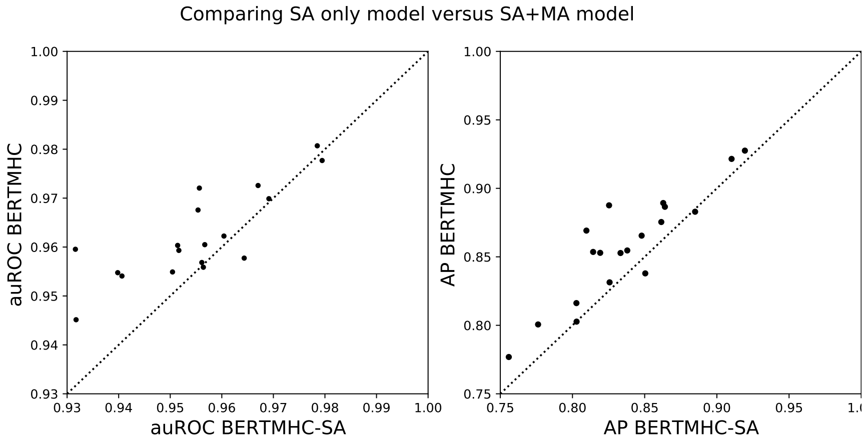
Table 3: Number of peptides and HLA type per patient

Patient	Number of peptides	HLA types
Patient1	6,999	DRB1*07:01, DRB1*11:01
Patient2	4,359	DRB1*15:02, DRB1*11:01
Patient3	5,053	DRB1*12:01, DRB1*11:01
Patient4	2,577	DRB1*12:01, DRB1*04:07
Patient5	3,570	DRB1*15:02, DRB1*11:01
Patient6	1,581	DRB1*04:07, DRB1*09:01

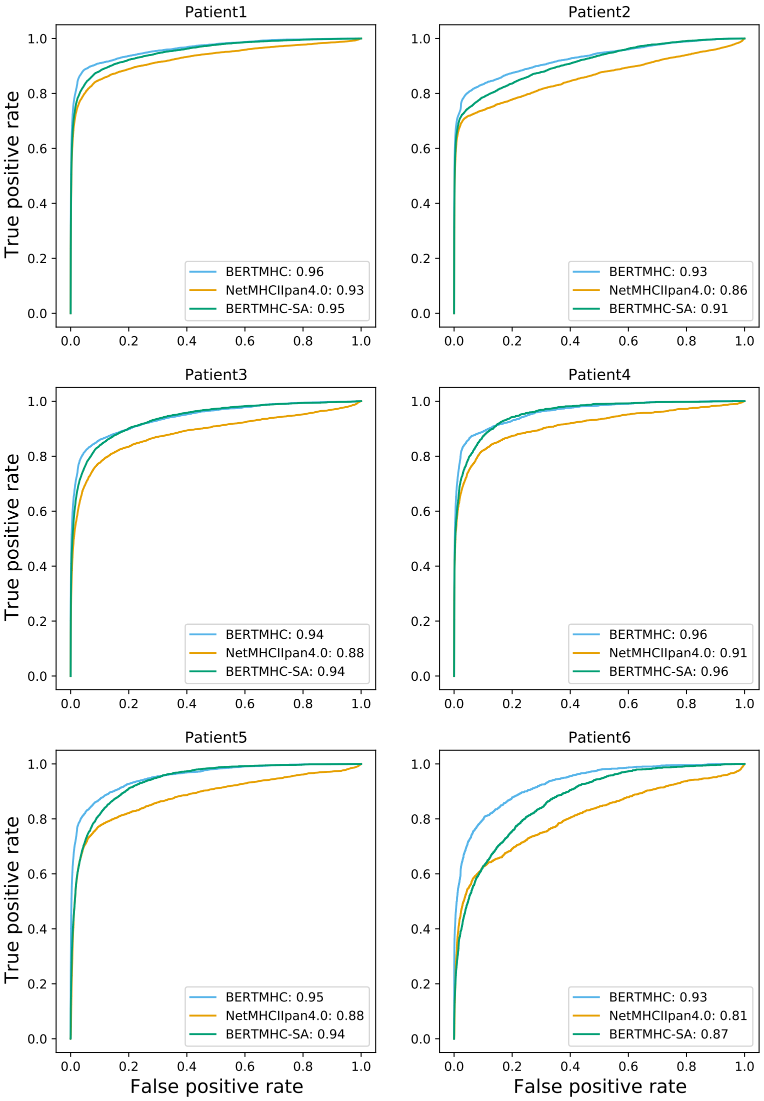
3 Supplementary Figures



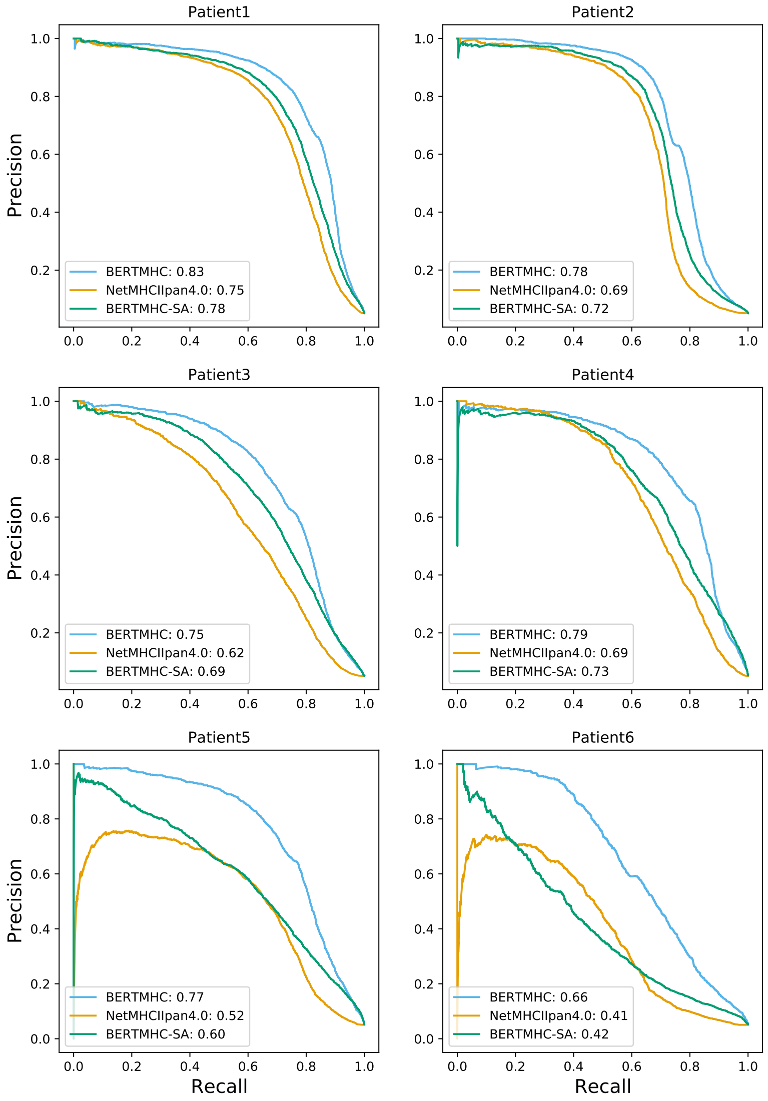
Supplementary Figure S1: Peptide length distribution for 6 patient samples.



Supplementary Figure S2: Comparing BERTMHC trained on both MA and SA data (y-axis) versus BERTMHC trained on SA data only (x-axis) in terms of auROC (left) and average precision (right). Each dot represent one allele in the SA data. Evaluation metrics for both models computed by performing out-of-fold predictions with the SA data.



Supplementary Figure S3: Comparing BERTMHC with NetMHCIIpan4.0 on patient mass spectrometry data. Receiver operating characteristic curve plotted for BERTMHC (cyan), BERTMHC-SA (green) and NetMHCIIpan4.0 (yellow)



Supplementary Figure S4: Comparing BERTMHC with NetMHCIIpan4.0 on patient mass spectrometry data. Precision-recall curve plotted for BERTMHC (cyan), BERTMHC-SA (green) and NetMHCIIpan4.0 (yellow)