Retrieval time prediction boosts confidence in identification of tumor-specific antigens in HLA immunopeptidomics

Eva Verzani¹, Karl R. Clauser¹, Susan Klaeger¹, Jennifer G. Abelin¹, Steven A. Carr¹

¹Broad Institute of MIT and Harvard, Cambridge, MA, USA

Tumor-specific peptide antigens presented on human leukocyte antigen (HLA) are a rich source of potential immunotherapy targets. While LC-MS/MS is now routinely used to sequence peptide antigens isolated in immunopeptidomics experiments, high-confidence identifications are desirable for advancing candidates to screening with immunogenicity assays that have limited scale. Confident identification of certain rare, high-value subpopulations of peptides, such as cancer neoantigens and disease-specific peptides derived from novel or unannotated open reading frames (nuORFs) is particularly challenging. Because the number of peptides in the high-value subpopulation is small, the no enzyme search space is large, and their abundance is often near the limits of sensitivity, false-positive identifications are more common. One orthogonal validation step in peptide identification is to predict retention time (RT) based on properties of the peptide and compare the alignment of predicted RT and observed RT; peptides with close alignment are more likely to be correct identifications. To accomplish this, we have applied deep learning–based retention time prediction algorithms (DeepLC and AutoRT) as an extra layer of validation on peptides of interest, including neoantigens, nuORFs, short open reading frames (smORFs), retained introns, and phosphorylated peptides. Given the variability of HLA peptides present between and within different immunopeptidome samples due to HLA allele diversity, it’s appropriate to use an algorithm that can be trained or calibrated using locally generated data covering a broad range of HLA alleles with unique peptide binding characteristics. After aggregate FDR thresholding of an immunopeptidomics dataset with Spectrum Mill, predicted RT plots typically show a greater correlation between predicted RT and observed RT for canonical peptide populations than for the noncanonical subsets mentioned above, indicating the necessity for further subset-specific FDR thresholding of these peptide populations. Higher confidence identification of peptides belonging to rare subsets is key to success in identifying immunogenic tumor-specific antigens and visualizing the landscape of a diseased cell’s immunopeptidome.

100321, https://doi.org/10.1016/j.mcpro.2022.100465