METHOD AND SYSTEMS FOR PREDICTION OF HLA CLASS II-SPECIFIC EPITOPES AND CHARACTERIZATION OF CD4+ T CELLS

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Application No. 62/783,914, filed on December 21, 2018; U.S. Provisional Application No. 62/826,827, filed on March 29, 2019; U.S. Provisional Application No. 62/855,379, filed on May 31, 2019; and U.S. Provisional Application No. 62/891,101, filed on August 23, 2019; each of which is incorporated herein by reference in its entirety.

SEQUENCE LISTING

[0001a] The instant application contains a Sequence Listing which has been submitted electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on January 31, 2020, is named 50401-735_601_SL.txt and is 27,306 bytes in size.

BACKGROUND

[0002] The major histocompatibility complex (MHC) is a gene complex encoding human leukocyte antigen (HLA) genes. HLA genes are expressed as protein heterodimers that are displayed on the surface of human cells to circulating T cells. HLA genes are highly polymorphic, allowing them to fine-tune the adaptive immune system. Adaptive immune responses rely, in part, on the ability of T cells to identify and eliminate cells that display disease-associated peptide antigens bound to human leukocyte antigen (HLA) heterodimers.

[0003] In humans, endogenous and exogenous proteins can be processed into peptides by the proteasome and by cytosolic and endosomal/lysosomal proteases and peptidases and presented by two classes of cell surface proteins encoded by MHC genes. These cell surface proteins are referred to as human leukocyte antigens (HLA class I and class II), and the group of peptides that bind them and elicit immune responses are termed HLA epitopes. HLA epitopes are a key component that enables the immune system to detect danger signals, such as pathogen infection and transformation of self. CD4+ T cells recognize class II MHC (HLA-DR, HLA-DQ, and HLA-DP) epitopes displayed on antigen presenting cells (APCs), such as dendritic cells and macrophages. The endogenous processing and presentation of HLA class II-ligands is a complex procedure and involves a variety of chaperones and a subset of enzymes that are not all well characterized. HLA class II-peptide presentation activates helper T cells, subsequently promoting B cell differentiation and antibody production as well as CTL responses. Activated helper T cells also secrete cytokines and chemokines that activate and induce differentiation of other T cells.

[0004] Understanding the peptide-binding preferences of every HLA class II heterodimer is the key to successfully predicting which cancer or tumor-specific antigens are likely to elicit the cancer or tumor-specific T cell responses. There is a need for methods of identifying and isolating specific HLA class II-associated peptides (e.g., neoantigen peptides). Such methodology and isolated molecules are useful, e.g., for the development of therapeutics, including but not limited to, immune based therapeutics.

SUMMARY

[0004a] In one aspect, the present disclosure refers to a method comprising:

(a) identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying the peptide sequence comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject;

(b) selecting a protein encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at most 50% and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence;

(c) contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; and

(d) optionally, purifying the identified peptide.

[0004b] In another aspect, the present disclosure refers to a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:

(a) obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells;

(c) determining or predicting that each of the plurality of peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions.

[0004c] In yet another aspect, the present disclosure refers to a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:

(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a

probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data, wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry;

training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and

(b) determining or predicting that each of the peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the set of presentation predictions.

[0004d] In yet another aspect, the present disclosure refers to a use of a composition in the manufacture of a medicament for the treatment of an autoimmune disease or condition in a subject in need thereof, the composition comprising a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising: (i) an epitope of an expressed protein identified or predicted to be presented by a class I or II MHC of a cell of the subject, and (ii) the class I or II MHC, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.

[0005] The methods and compositions described herein find uses in a wide range of applications. For example, the methods and compositions described herein can be used to identify immunogenic antigen

performance of a prediction model; identifying allele-specific HLA class II binding cores from HLA-ligandome LC-MS/MS datasets; utilizing machine learning algorithms to improve HLA class II-ligand and epitope prediction; and/or identifying biological variables that impact HLA class II -ligand presentation and improve HLA class II epitope prediction, such as gene expression, cleavability, gene bias, cellular localization, and secondary structure.

[0012] Provided herein is a method comprising: (a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells; and (b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject; wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 according to a presentation PPV determination method.

[0013] Provided herein is a method comprising: (a) processing amino acid information of a plurality of peptide sequences of encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences, wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or an HLA class II protein analog; and (b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject; wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1 according to a binding PPV determination method.

[0014] In some examples, the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells.

[0015] In some examples, the method comprises ranking, based on the presentation predictions, at least two peptides identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0016] In some examples, the method comprises selecting one or more peptides of the two or more ranked peptides.

[0017] In some examples, the method comprises selecting one or more peptides of the plurality that were identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0018] In some examples, the method comprises selecting one or more peptides of two or more peptides ranked based on the presentation predictions.

[0019] In some examples, the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test presentation predictions, each test presentation prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 500 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 499 decoy peptide sequences contained within a protein encoded by a genome of an organism, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of 1:499 of the at least one hit peptide sequence to the at least 499 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to be presented by the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

[0020] In some examples, the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 20 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising at least one peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, such as a single HLA protein expressed in cells (e.g., mono-allelic cells), wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to bind to the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

[0021] In some examples, no amino acid sequence overlap exist among the at least one hit peptide sequence and the decoy peptide sequences.

[0022] In some examples, the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98 or 0.99.

[0023] In some examples, the at least one hit peptide sequence comprises at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100 hit peptide sequences.

[0024] In some examples, the at least 499 decoy peptide sequences comprises at least 500 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000, 5100, 5200, 5300, 5400, 5500, 5600, 5700, 5800, 5900, 6000, 6100, 6200, 6300, 6400, 6500, 6600, 6700, 6800, 6900, 7000, 7100, 7200, 7300, 7400, 7500, 7600, 7700, 7800, 7900, 8000, 8100, 8200, 8300, 8400, 8500, 8600, 8700, 8800, 8900, 9000, 9100, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000, 21000, 22000, 23000, 24000, 25000, 26000, 27000, 28000, 29000, 30000, 31000, 32000, 33000, 34000, 35000, 36000, 37000, 38000, 39000, 40000, 41000, 42000, 43000, 44000, 45000, 46000, 47000, 48000, 49000, 50000, 52500, 55000, 57500, 60000, 62500, 65000, 67500, 70000, 72500, 75000, 77500, 80000, 82500, 85000, 87500, 90000, 92500, 95000, 97500, 100000, 125000, 150000, 175000, 200000, 225000, 250000, 275000, 300000, 325000, 350000, 375000, 400000, 425000, 450000, 475000, 500000, 600000, 700000, 800000, 900000 or 1000000 decoy peptide sequences. One of skill in the art is able to recognize that changing the ratio of hit : decoy changes the PPV.

[0025] In some examples, the at least 500 test peptide sequences comprises at least 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000, 5100, 5200, 5300, 5400, 5500, 5600, 5700, 5800, 5900, 6000, 6100, 6200, 6300, 6400, 6500, 6600, 6700, 6800, 6900, 7000, 7100, 7200, 7300, 7400, 7500, 7600, 7700, 7800, 7900, 8000, 8100, 8200, 8300, 8400, 8500, 8600, 8700, 8800, 8900, 9000, 9100, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000, 21000, 22000, 23000, 24000, 25000, 26000, 27000, 28000, 29000, 30000, 31000, 32000, 33000, 34000, 35000, 36000, 37000, 38000, 39000, 40000, 41000, 42000, 43000, 44000, 45000, 46000, 47000, 48000, 49000, 50000, 52500, 55000, 57500, 60000, 62500, 65000, 67500, 70000, 72500, 75000, 77500, 80000, 82500, 85000, 87500, 90000, 92500, 95000, 97500, 100000, 125000, 150000, 175000, 200000, 225000, 250000, 275000, 300000, 325000, 350000, 375000, 400000, 425000, 450000, 475000, 500000, 600000, 700000, 800000, 900000 or 1000000 test peptide sequences.

[0026] In some examples, the top percentage is a top 0.20%, 0.30%, 0.40%, 0.50%, 0.60%, 0.70%, 0.80%, 0.90%, 1.00%, 1.10%, 1.20%, 1.30%, 1.40%, 1.50%, 1.60%, 1.70%, 1.80%, 1.90%, 2.00%, 2.10%, 2.20%, 2.30%, 2.40%, 2.50%, 2.60%, 2.70%, 2.80%, 2.90%, 3.00%, 3.10%, 3.20%, 3.30%, 3.40%, 3.50%, 3.60%, 3.70%, 3.80%, 3.90%, 4.00%, 4.10%, 4.20%, 4.30%, 4.40%, 4.50%, 4.60%, 4.70%, 4.80%, 4.90%, 5.00%, 5.10%, 5.20%, 5.30%, 5.40%, 5.50%, 5.60%, 5.70%, 5.80%, 5.90%, 6.00%, 6.10%, 6.20%, 6.30%, 6.40%, 6.50%, 6.60%, 6.70%, 6.80%, 6.90%, 7.00%, 7.10%, 7.20%, 7.30%, 7.40%, 7.50%, 7.60%, 7.70%, 7.80%, 7.90%, 8.00%, 8.10%, 8.20%, 8.30%, 8.40%, 8.50%, 8.60%, 8.70%, 8.80%, 8.90%, 9.00%, 9.10%, 9.20%, 9.30%, 9.40%, 9.50%, 9.60%, 9.70%, 9.80%, 9.90%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19% or 20%.

[0027] In some examples, the at least one hit peptide sequence comprises at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 or 100 hit peptide sequences.

[0028] In some examples, the at least 19 decoy peptide sequences comprises at least 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000, 5100, 5200, 5300, 5400, 5500, 5600, 5700, 5800, 5900, 6000, 6100, 6200, 6300, 6400, 6500, 6600, 6700, 6800, 6900, 7000, 7100, 7200, 7300, 7400, 7500, 7600, 7700, 7800, 7900, 8000, 8100, 8200, 8300, 8400, 8500, 8600, 8700, 8800, 8900, 9000, 9100, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000, 21000, 22000, 23000, 24000, 25000, 26000, 27000, 28000, 29000, 30000, 31000, 32000, 33000, 34000, 35000, 36000, 37000, 38000, 39000, 40000, 41000, 42000, 43000, 44000, 45000, 46000, 47000, 48000, 49000, 50000, 52500, 55000, 57500, 60000, 62500, 65000, 67500, 70000, 72500, 75000, 77500, 80000, 82500, 85000, 87500, 90000, 92500, 95000, 97500, 100000, 125000, 150000, 175000, 200000, 225000, 250000, 275000, 300000, 325000, 350000, 375000, 400000, 425000, 450000, 475000, 500000, 600000, 700000, 800000, 900000 or 1000000 decoy peptide sequences.

[0029] In some examples, the at least 20 test peptide sequences comprises at least wherein the at least 500 test peptide sequences comprises at least 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000, 5100, 5200, 5300, 5400, 5500, 5600, 5700, 5800, 5900, 6000, 6100, 6200, 6300, 6400, 6500, 6600, 6700, 6800, 6900, 7000, 7100, 7200, 7300, 7400, 7500, 7600, 7700, 7800, 7900, 8000, 8100, 8200, 8300, 8400, 8500, 8600, 8700, 8800,

8900, 9000, 9100, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000, 21000, 22000, 23000, 24000, 25000, 26000, 27000, 28000, 29000, 30000, 31000, 32000, 33000, 34000, 35000, 36000, 37000, 38000, 39000, 40000, 41000, 42000, 43000, 44000, 45000, 46000, 47000, 48000, 49000, 50000, 52500, 55000, 57500, 60000, 62500, 65000, 67500, 70000, 72500, 75000, 77500, 80000, 82500, 85000, 87500, 90000, 92500, 95000, 97500, 100000, 125000, 150000, 175000, 200000, 225000, 250000, 275000, 300000, 325000, 350000, 375000, 400000, 425000, 450000, 475000, 500000, 600000, 700000, 800000, 900000 or 1000000 test peptide sequences test peptide sequences.

[0030] In some examples, the top percentage is a top 5%, 6%, 7%, 8%, 9%, 10%, 11%, 12%, 13%, 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, or 40%.

[0031] In some examples, the PPV is greater than the respective PPV of column 2 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11. In some examples, the PPV is at least equal to the respective PPV of column 3 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11.

[0032] In some examples, the PPV is equal to or greater than the respective PPV of column 2 of Table 12 for the protein encoded by an HLA class II allele.

[0033] In some examples, the PPV is greater than the respective PPV of column 2 of Table 16 for the protein encoded by an HLA class II allele.

[0034] In some examples, the subject is a single subject.

[0035] In some examples, the subject is a mammal.

[0036] In some examples, the subject is a human.

[0037] In some examples, the training cells are cells expressing a single protein encoded by a class II HLA allele of a cell of the subject.

[0038] In some examples, the training cells are monoallelic HLA cells, or cells expressing an HLA allele with an affinity tag.

[0039] In some examples, the cell of the subject comprises cancer cells.

[0040] In some examples, the method is for identifying peptide sequences.

[0041] In some examples, the method is for selecting peptide sequences.

[0042] In some examples, the method is for preparing a cancer therapy.

[0043] In some examples, the method is for preparing a subject-specific cancer therapy.

[0044] In some examples, the method is for preparing a cancer cell-specific cancer therapy.

[0045] In some examples, each peptide sequence of the plurality of peptide sequences is associated with a cancer.

[0046] In some examples, at least one peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

[0047] In some examples, each peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.

[0048] In some examples, at least one peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

[0049] In some examples, each peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.

[0050] In some examples, each peptide sequence of the plurality of peptide sequences is expressed by a cancer cell of the subject.

[0051] In some examples, at least one peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

[0052] In some examples, each peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.

[0053] In some examples, at least one peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject

[0054] In some examples, each peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.

[0055] In some examples, the method comprises obtaining the plurality of peptide sequences of the subject.

[0056] In some examples, the method comprises obtaining a plurality of polynucleotide sequences of the subject.

[0057] In some examples, the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject, or by a pathogen or virus in the subject.

[0058] In some examples, the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject by a computer processor.

[0059] In some examples, the method comprises obtaining a plurality of polynucleotide sequences of the subject by genomic or exomic sequencing.

[0060] In some examples, the method comprises obtaining a plurality of polynucleotide sequences of the subject by whole genome sequencing or whole exome sequencing.

[0061] In some examples, processing comprises processing by a computer processor

[0062] In some examples, processing comprises generating a plurality of predictor variables based at least on the amino acid information of the plurality of peptide sequences

[0063] In some examples, processing the plurality of predictor variables using the machine-learning HLA-peptide presentation prediction model.

[0064] In some examples, the that one or more proteins encoded by a class II HLA allele of a cell of the subject are one or more proteins encoded by a class II HLA allele that are expressed by the subject.

[0065] In some examples, the that one or more proteins encoded by a class II HLA allele of a cell of the subject are one or more proteins encoded by a class II HLA allele that are expressed by cancer cells of the subject.

[0066] In some examples, the that one or more proteins encoded by a class II HLA allele of a cell of the subject is a single protein encoded by a class II HLA allele of a cell of the subject.

[0067] In some examples, the that one or more proteins encoded by a class II HLA allele of a cell of the subject is two, three, four, five or six or more proteins encoded by a class II HLA allele of a cell of the subject.

[0068] In some examples, the that one or more proteins encoded by a class II HLA allele of a cell of the subject is each protein encoded by a class II HLA allele of a cell of the subject.

[0069] In some examples, the method further comprises administering to the subject a composition comprising one or more of the selected sub-set of peptide sequences.

[0070] In some examples, identifying the plurality of peptide sequences comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject, wherein each of the plurality of the peptides comprise at least one mutation, which is present in the cancer cell of the subject, and not present in the normal cell of the subject.

[0071] In some examples, the machine-learning HLA-peptide presentation prediction model comprises a plurality of predictor variables identified at least based on the training data, wherein the training data comprises training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information and the presentation likelihood generated as output based on the amino acid position information and the plurality of predictor variables.

[0072] In some examples, identifying comprises identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold presentation prediction probability value of being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0073] In some examples, one or more of the 0.2% of the plurality of test peptide sequences predicted to be presented by the by the machine learning HLA peptide presentation prediction model has a probability greater than the threshold presentation prediction probability value of being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0074] In some examples, each of the 0.2% of the plurality of test peptide sequences predicted to be presented by the by the machine learning HLA peptide presentation prediction model has a probability greater than the threshold presentation prediction probability value of being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0075] In some examples, the number of positives is constrained to be equal to the number of hits.

[0076] In some examples, the mass spectrometry is mono-allelic mass spectrometry.

[0077] In some examples, the peptides are presented by a HLA protein expressed in cells through autophagy.

[0078] In some examples, the peptides are presented by a HLA protein expressed in cells through phagocytosis.

[0079] In some examples, the plurality of predictor variables comprises expression level predictor of the source protein comprising the peptide.

[0080] In some examples, the plurality of predictor variables comprises stability predictor of the source protein comprising the peptide.

[0081] In some examples, the plurality of predictor variables comprises degradation rate predictor of the source protein comprising the peptide.

[0082] In some examples, the plurality of predictor variables comprises protein cleavability predictor of the source protein comprising the peptide.

[0083] In some examples, the plurality of predictor variables comprises cellular or tissue localization predictor of the source protein comprising the peptide.

[0084] In some examples, the plurality of predictor variables comprises a predictor for the intracellular processing mode of the source protein comprising the peptide, wherein processing mode of the source protein comprises predictor for whether the source protein is subject to autophagy, phagocytosis, and intracellular transport, among others.

[0085] In some examples, quality of the training data is increased by using a plurality of quality metrics.

[0086] In some examples, the plurality of quality metrics comprises common contaminant peptide removal, high scored peak intensity, high score, and high mass accuracy.

[0087] In some examples, a scored peak intensity is at least 50%.

[0088] In some examples, the scored peak intensity is at least 60%.

[0089] In some examples, a score is at least 7.

[0090] In some vs, a mass accuracy is at most 5 ppm.

[0091] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single immunoprecipitated HLA protein expressed in cells.

[0092] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single exogenous HLA protein expressed in cells.

[0093] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single recombinant HLA protein expressed in cells.

[0094] In some examples, the plurality of predictor variables comprises a peptide-HLA affinity predictor variable.

[0095] In some examples, the peptides presented by the HLA protein comprise peptides identified by searching a no-enzyme specificity without modification peptide database.

[0096] In some examples, the peptides presented by the HLA protein comprise peptides identified by searching a peptide database using a reversed-database search strategy.

[0097] In some examples, the HLA protein comprises an HLA-DR, HLA-DQ, or an HLA-DP protein.

[0098] In some examples, the HLA protein comprises an HLA class II protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-

DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01, HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02, HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, HLA-DRB5*01:01.

[0099] In some examples, the HLA-DR is paired with paired with DRA*01:01.

[0100] In some examples, the HLA protein is a HLA class II protein selected from the group consisting of: DPA*01:03/DPB*04:01, DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*04:01, DRB1*04:02, DRB1*04:04, DRB1*04:05, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*09:01, DRB1*11:01, DRB1*11:02, DRB1*11:04, DRB1*12:01, DRB1*13:01, DRB1*13:02, DRB1*13:03, DRB1*14:01, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*16:02, DRB3*01:01, DRB3*02:01, DRB3*02:02, DRB3*03:01, DRB4*01:01, DRB4*01:03 and DRB5*01:01.

[0101] In some examples, the HLA-DR protein comprises a DRA*01:01 in the dimer.

[0102] In some examples, the HLA protein comprises an HLA-DP protein selected from the group consisting of: DPB1*01:01, DPB1*02:01, DPB1*02:02, DPB1*03:01, DPB1*04:01, DPB1*04:02, DPB1*05:01, DPB1*06:01, DPB1*11:01, DPB1*13:01, DPB1*17:01.

[0103] In some examples, the HLA-DP protein is paired comprising DPA1*01:03.

[0104] In some examples, the HLA protein comprises an HLA-DQ protein complex selected from the group consisting of: A1*01:01+B1*05:01, A1*01:02+B1*06:02, A1*01:02+B1*06:04, A1*01:03+B1*06:03, A1*02:01+B1*02:02, A1*02:01+B1*03:03, A1*03:01+B1*03:02, A1*03:03+B1*03:01, A1*05:01+B1*02:01 and A1*05:05+B1*03:01.

[0105] In some examples, the peptides presented by the HLA protein comprise peptides identified by comparing a MS/MS spectra of the HLA-peptides with MS/MS spectra of one or more peptides or proteins in a peptide or protein database.

[0106] In some examples, the mutation is selected from the group consisting of a point mutation, a splice site mutation, a frameshift mutation, a read-through mutation, and a gene fusion mutation.

[0107] In some examples, the peptides presented by the HLA protein have a length of from 15-40 amino acids.

[0108] In some examples, the peptides presented by the HLA protein comprise peptides identified by identifying peptides presented by an HLA protein by comparing a MS/MS spectra of the HLA-peptides with MS/MS spectra of one or more peptides or proteins in a peptide or protein database.

[0109] In some examples, the personalized cancer therapy further comprises an adjuvant.

[0110] In some examples, the personalized cancer therapy further comprises an immune checkpoint inhibitor.

[0111] In some examples, the training data comprises structured data, time-series data, unstructured data, relational data, or any combination thereof.

[0112] In some examples, the unstructured data comprises image data.

[0113] In some examples, the relational data comprises data from a customer system, an enterprise system, an operational system, a website, web accessible application program interface (API), or any combination thereof.

[0114] In some examples, the training data is uploaded to a cloud-based database.

[0115] In some examples, the training is performed using convolutional neural networks.

[0116] In some examples, the convolutional neural networks comprise at least two convolutional layers.

[0117] In some examples, the convolutional neural networks comprise at least one batch normalization step.

[0118] In some examples, the convolutional neural networks comprise at least one spatial dropout step.

[0119] In some examples, the convolutional neural networks comprise at least one global max pooling step.

[0120] In some examples, the convolutional neural networks comprise at least one dense layer.

[0121] In some examples, identifying peptide sequences comprises identifying peptide sequences with a mutation expressed in cancer cells of a subject.

[0122] In some examples, identifying peptide sequences comprises identifying peptide sequences not expressed in normal cells of a subject.

[0123] In some examples, identifying peptide sequences comprises identifying viral peptide sequences.

[0124] In some examples, identifying peptide sequences comprises identifying overexpressed peptide sequences.

[0125] Provided herein is a method for identifying HLA class II specific peptides for immunotherapy for a subject, comprising: obtaining, by a computer processor, a candidate peptide comprising an epitope, and a plurality of peptide sequences, each comprising the epitope; processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences to an immune cell, each presentation prediction indicative of a likelihood that one or more proteins encoded by an HLA class II allele can present a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; selecting a protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject, predicted to bind to the candidate peptide by the machine-learning HLA-peptide presentation prediction model, wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the candidate peptide to an immune cell; contacting the candidate peptide with the selected protein,

such that the candidate peptide competes with a placeholder peptide associated with the selected protein; and identifying the candidate peptide as a peptide for immunotherapy specific for the selected protein based on whether the candidate peptide displaces the placeholder

[0126] In some examples, obtaining comprises identifying the candidate peptide, wherein identifying the candidate peptide comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject.

[0127] In some examples, processing comprises identifying a plurality of predictor variables based at least on the amino acid information of the plurality of peptide sequences, and processing the plurality of predictor variables using the machine-learning HLA-peptide presentation prediction model.

[0128] In some examples, the machine-learning HLA-peptide presentation prediction model comprises a plurality of predictor variables identified at least based on the training data, wherein the training data comprises: training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information and the presentation likelihood generated as output based on the amino acid position information and the plurality of predictor variables.

[0129] In some examples, the number of positives is constrained to be equal to the number of hits.

[0130] In some examples, the mass spectrometry is mono-allelic mass spectrometry.

[0131] In some examples, the plurality of predictor variables comprises any one or more of: expression level predictor, stability predictor, degradation rate predictor, cleavability predictor, cellular or tissue localization predictor, and intracellular processing mode comprising autophagy, phagocytosis, and intracellular transport predictor, of the source protein comprising the peptide.

[0132] In some examples, quality of the training data is increased by using a plurality of quality metrics.

[0133] In some examples, the plurality of quality metrics comprises common contaminant peptide removal, high scored peak intensity, high score, and high mass accuracy.

[0134] In some examples, a scored peak intensity is at least 50%.

[0135] In some examples, the scored peak intensity is at least 60%.

[0136] In some examples, the placeholder peptide is a CLIP peptide.

[0137] In some examples, the placeholder peptide is a CMV peptide.

[0138] In some examples, the3 method further comprises measuring the IC50 of displacement of the placeholder peptide by the target peptide.

[0139] In some examples, the IC50 of displacement of the placeholder peptide by the target peptide is less than 500 nM.

[0140] In some examples, the at least one protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject is an HLA class II tetramer or multimer.

[0141] In some examples, the target peptide is further identified by mass spectrometry.

[0142] In some examples, the at least one protein encoded by the HLA class II allele of a cell of the subject is a recombinant protein.

[0143] In some examples, the at least one protein encoded by the HLA class II allele of a cell of the subject is expressed in a eukaryotic cell.

[0144] In some examples, the peptides are presented by a HLA protein expressed in cells through autophagy.

[0145] In some examples, the peptides are presented by a HLA protein expressed in cells through phagocytosis.

[0146] In some examples, the peptides presented by a HLA protein expressed in cells are peptides presented by a single immunoprecipitated HLA protein expressed in cells.

[0147] In some examples, the peptides presented by a HLA protein expressed in cells are peptides presented by a single exogenous HLA protein expressed in cells.

[0148] In some examples, the peptides presented by a HLA protein expressed in cells are peptides presented by a single recombinant HLA protein expressed in cells.

[0149] In some examples, the plurality of predictor variables comprises a peptide-HLA affinity predictor variable.

[0150] In some examples, the peptides presented by the HLA protein comprise peptides identified by searching a no-enzyme specificity without modification peptide database.

[0151] In some examples, the peptides presented by the HLA protein comprise peptides identified by searching a peptide database using a reversed-database search strategy.

[0152] In some examples, the HLA protein comprises an HLA-DR, HLA-DQ, or an HLA-DP protein.

[0153] In some embodiments, the immunotherapy is cancer immunotherapy.

[0154] In some embodiments, the epitope is a cancer specific epitope.

[0155] In some embodiments, the at least one protein encoded by the HLA class II allele comprises at least an alpha 1 subunit and a beta 1 subunit of the HLA protein, present in dimer form.

[0156] In some examples, the identity of the peptide is known.

[0157] In some examples, the identity of the peptide is not known.

[0158] In some examples, the identity of the peptide is determined by mass spectrometry.

[0159] In some examples, peptide exchange assay comprises detection of peptide fluorescent probes or tags.

[0160] In some examples, in the placeholder peptide is a CLIP peptide. In some examples, the placeholder peptide has an amino acid sequence of PVSKMRMATPLLMQA (SEQ ID NO: 1).

[0161] In some examples, the polynucleic acid construct comprises an expression vector, further comprising one or more of: a promoter, a secretion signal, dimerization factors, ribosomal skipping sequence, one or more tags for purification and/or detection.

[0162] In some examples, the placeholder peptide sequence is encoded by a nucleic acid sequence within the vector.

[0163] In some examples, a sequence encoding a cleavable domain is placed in between the sequence encoding the placeholder peptide and the HLA beta1 peptide.

[0164] Provided herein is a method for assaying immunogenicity of a MHC class II binding peptide, comprising: selecting a protein encoded by an HLA class II allele predicted by a machine-learning HLA-peptide presentation prediction model to bind to the MHC class II binding peptide, wherein the machine-learning HLA-peptide presentation prediction model is configured to generate a presentation prediction for a given peptide sequence, the presentation prediction indicative of a likelihood that one or more proteins encoded by the HLA class II allele can present the given peptide sequence, and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the MHC class II binding peptide; contacting the peptide with the selected protein such that the peptide competes with a placeholder peptide associated with the selected protein, and displaces the placeholder peptide, thereby forming a complex comprising the HLA class II protein and the MHC class II binding peptide; contacting the complex with a CD4+ T cell, and assaying for one or more of activation parameters of the CD4+T cell, selected from the group consisting of: induction of a cytokine, induction of a chemokine, and expression of a cell surface marker.

[0165] In some examples, the HLA class II allele is a tetramer or multimer.

[0166] In some examples, the cytokine is IL-2.

[0167] Provided herein is a method for inducing a CD4+ T cells activation in a subject for cancer immunotherapy, the method comprising: identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying the peptide sequence comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject; selecting a protein encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at most 50%.and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence; contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; optionally, purifying the identified peptide; and administering an effective amount of a polypeptide comprising a sequence of the identified peptide or a polynucleotide encoding the polypeptide to the subject.

[0168] Provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising: obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence; processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells; determining or predicting that each of the plurality of peptide sequences of

the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions; and administering to the subject a composition comprising the drug.

[0169] Provided herein is a method for manufacturing HLA class II tetramers or multimers by conjugation of four individual HLA protein alpha1 and beta1 heterodimers, the method comprising: expressing in a eukaryotic cell, a vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, wherein the heterodimer is associated with a placeholder peptide, purifying the secreted heterodimer from cell medium, validating the peptide binding activity using peptide exchange assay, adding streptavidin thereby conjugating heterodimers into tetramers, purifying the tetramers and having a yield of greater than 1 mg/L. Multimers, for example pentamers, hexamers or octamers can also be likewise generated, which are equally contemplated herein.

[0170] In some examples, the vector comprises a CMV promoter.

[0171] In some examples, the vector comprises a sequence encoding a placeholder peptide linked via a cleavable site to the beta 1 chain.

[0172] In some examples, peptide exchange assay involves prior cleavage of the placeholder peptide from the beta chain.

[0173] In some examples, the cleavable site is a thrombin cleavage site.

[0174] In some examples, peptide exchange assay is a FRET assay.

[0175] In some examples, the purification is by any one of: column chromatography, ion exchange chromatography, size exclusion chromatography, affinity chromatography, or LC-MS.

[0176] Provided herein is an HLA class II tetramer or multimer comprising either HLA-DR, or HLA-DP, or HLA-DQ heterodimers, each heterodimer comprising an alpha and a beta chain, wherein the heterodimer is purified and present at a concentration of greater than 1 mg/L.

[0177] In some examples, the HLA class II tetramers are selected from Table 8A-8C.

[0178] In some examples, the HLA class II tetramer comprises heterodimer pairs selected from the group consisting of: an HLA-DR, an HLA-DP, and an HLA-DQ protein.

[0179] In some examples, the HLA protein is an HLA class II protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01,HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02,HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-

DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, and HLA-DRB5*01:01.

[0180] In some examples, the heterodimer pair is expressed in a eukaryotic cell.

[0181] In some examples, the heterodimer pairs are encoded by a vector.

[0182] Provided herein is a vector, wherein the vector comprises a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein described herein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, wherein the secreted heterodimer is optionally associated with a placeholder peptide.

[0183] Provided herein is a cell, comprising a vector described herein.

[0184] In some examples, the HLA class II heterodimers are secreted from eukaryotic cells into cell culture medium, which is further purified by any one of: column chromatography, ion exchange chromatography, size exclusion chromatography, affinity chromatography or LC-MS.

[0185] Provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising: obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence; processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; and determining or predicting that at least one of the plurality of peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the plurality of presentation predictions.

[0186] Provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising: inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data, wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; determining or predicting that each of the peptide sequences of the polypeptide sequence

would not be immunogenic to the subject based on the set of presentation predictions; and administering to the subject a composition comprising the drug.

[0187] Provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising: inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data; wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; determining or predicting that at least one of the peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the set of presentation predictions.

[0188] Provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising: obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence; processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells; determining or predicting that each of the plurality of peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions; and administering to the subject a composition comprising the drug.

[0189] In some examples, the method further comprises deciding not to administer the drug to the subject.

[0190] In some examples, the drug comprises an antibody or binding fragment thereof.

[0191] In some examples, the peptide sequences of the polypeptide sequence have a length of 8, 9, 10, 11, or 12 amino acids, and wherein the protein encoded by a class I or II MHC allele of a cell of the subject is a protein encoded by a class I MHC allele of a cell of the subject.

[0192] In some examples, the peptide sequences of the polypeptide sequence have a length of 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 amino acids, and wherein the protein encoded by a class I or II MHC allele of a cell of the subject is a protein encoded by a class II MHC allele of a cell of the subject.

[0193] Provided herein is a method of treating a subject with an autoimmune disease or condition comprising: (a) identifying or predicting an epitope of an expressed protein presented by a class I or II MHC of a cell of the

subject, wherein a complex comprising the identified or predicted epitope and the class I or II MHC is targeted by a CD8 or CD4 T cell of the subject; (b) identifying a T cell receptor (TCR) that binds to the complex; (c) expressing the TCR in a regulatory T cell from the subject or an allogeneic regulatory T cell; and (d) administering the regulatory T cell expressing the TCR to the subject.

[0194] In some examples, the autoimmune disease or condition is diabetes.

[0195] In some examples, the cell is an islet cell.

[0196] Provided herein is a method of treating a subject with an autoimmune disease or condition, comprising administering to the subject a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising: (i) an epitope of an expressed protein identified or predicted to be presented by a class I or II MHC of a cell of the subject, and (ii) the class I or II MHC, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.

[0197] Provided herein is a computer system for identifying peptide sequences for a personalized cancer therapy of a subject, comprising: a database that is configured to store a plurality of peptide sequences of the subject; and one or more computer processors operatively coupled to said database, wherein said one or more computer processors are individually collectively programmed to: process amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class II MHC allele of a cell of the subject can present a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; and select a subset of the plurality of peptide sequences for the personalized cancer therapy of the subject based at least on the plurality of presentation predictions.

[0198] Provided herein is a computer system for identifying HLA class II specific peptides for immunotherapy for a subject, comprising: a database that is configured to store a candidate peptide comprising an epitope, and a plurality of peptide sequences, each comprising the epitope; and one or more computer processors operatively coupled to said database, wherein said one or more computer processors are individually collectively programmed to: process amino acid information of the plurality of peptide sequences a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences to an immune cell, each presentation prediction indicative of a likelihood that one or more proteins encoded by an HLA class II allele can present a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; select a protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject, predicted to bind to the candidate peptide by the machine-learning HLA-peptide presentation prediction model, wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the candidate peptide to an immune cell; and identify the candidate

subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences, wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or an HLA class II protein analog; and identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject; wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 50 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising a peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and 5% of the plurality of test peptide sequences are predicted to bind to the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

[0207] In some examples, the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells

[0208] In some examples, one or more of the 0.2% of the plurality of test peptide sequences predicted to be presented by the by the machine learning HLA peptide presentation prediction model has a probability greater than the threshold presentation prediction probability value of being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0209] In some examples, each of the 0.2% of the plurality of test peptide sequences predicted to be presented by the by the machine learning HLA peptide presentation prediction model has a probability greater than the threshold presentation prediction probability value of being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.

[0210] In some examples, the PPV is greater than the respective PPV of column 2 of Table 11 for the protein encoded by the corresponding HLA allele of Table 13. In some examples, the PPV is at least equal to the respective PPV of column 3 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11.

[0211] In some examples, the PPV is greater than the respective PPV of column 2 of Table 12 for the protein encoded by an HLA class II allele.

[0212] In some examples, the PPV is at least equal to the respective PPV of column 2 of Table 16 for the protein encoded by the corresponding HLA allele of Table 16.

[0213] Provided herein is a method for preparing a personalized cancer therapy, the method comprising: identifying peptide sequences, wherein the peptide sequences are associated with cancer, wherein identifying comprises comparing DNA, RNA or protein sequences from the cancer cells of the subject to DNA, RNA or protein sequences from the normal cells of the subject; inputting amino acid position information of the peptide sequences identified, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences identified, each presentation prediction representing a probability that one or more proteins encoded by an HLA class II allele of a cell of the subject will present a given sequence of a peptide sequence identified; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data wherein the training data comprises: sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and selecting a subset of the peptide sequences identified based on the set of presentation predictions for preparing the personalized cancer therapy; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50%.

[0214] Provided herein is a method comprising training a machine-learning HLA-peptide presentation prediction model, wherein training comprises inputting amino acid position information sequences of HLA-peptides isolated from one or more HLA-peptide complexes from a cell expressing an HLA class II allele into the HLA-peptide presentation prediction model using a computer processor; the machine-learning HLA-peptide presentation prediction model comprising: a plurality of predictor variables identified at least based on training data that comprises: sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information of training peptides, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and a presentation likelihood generated as output based on the amino acid position information and the predictor variables.

[0215] In some examples, the presentation model has a positive predictive value of at least 0.25 at a recall rate at least 0.1%, from 0.1%-50% or at the most 50%.

[0216] In some examples, the presentation model has a positive predictive value of at least 0.4 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50%.

[0217] In some examples, the presentation model has a positive predictive value of at least 0.6 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50%.

[0218] In some examples, the mass spectrometry is mono-allelic mass spectrometry.

[0219] In some examples, the peptides are presented by an HLA protein expressed in cells through autophagy.

[0220] In some examples, the peptides are presented by an HLA protein expressed in cells through phagocytosis.

[0221] In some examples, quality of the training data is increased by using a plurality of quality metrics.

[0222] In some examples, the plurality of quality metrics comprises common contaminant peptide removal, high scored peak intensity, high score, and high mass accuracy.

[0223] In some examples, the scored peak intensity is at least 50%.

[0224] In some examples, the scored peak intensity is at least 60%.

[0225] In some examples, a score is at least 7.

[0226] In some examples, a mass accuracy is at most 5 ppm.

[0227] In some examples, a mass accuracy is at most 2 ppm.

[0228] In some examples, a backbone cleavage score is at least 5.

[0229] In some examples, a backbone cleavage score is at least 8.

[0230] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single immunoprecipitated HLA protein expressed in cells.

[0231] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single exogenous HLA protein expressed in cells.

[0232] In some examples, the peptides presented by an HLA protein expressed in cells are peptides presented by a single recombinant HLA protein expressed in cells.

[0233] In some examples, the plurality of predictor variables comprises a peptide-HLA affinity predictor variable.

[0234] In some examples, the plurality of predictor variables comprises a source protein expression level predictor variable.

[0235] In some examples, the plurality of predictor variables comprises a peptide cleavability predictor variable.

[0236] In some examples, the training peptide sequence information comprises sequences from the peptides presented by the HLA protein, which comprise peptides identified by searching a no-enzyme specificity without modification to a peptide database. In some embodiments, the peptides presented by the HLA protein comprise peptides identified by searching the de novo peptide sequencing tools.

[0237] In some examples, the peptides presented by the HLA protein comprise peptides identified by searching a peptide database using a reversed-database search strategy.

[0238] In some examples, the HLA protein comprises an HLA-DR, and HLA-DP or an HLA-DQ protein. In some embodiments, the HLA protein comprises an HLA-DR protein selected from the group consisting of an HLA-DR, and HLA-DP or an HLA-DQ protein. In some embodiments, the HLA protein comprises an HLA-DR protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-

DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01,HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02,HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, and HLA-DRB5*01:01.

[0239] In some examples, the peptides presented by the HLA protein comprise peptides identified by comparing MS/MS spectra of the HLA-peptides with MS/MS spectra of one or more HLA-peptides in a peptide database.

[0240] In some examples, the mutation is selected from the group consisting of a point mutation, a splice site mutation, a frameshift mutation, a read-through mutation, and a gene fusion mutation.

[0241] In some examples, the peptides presented by the HLA protein have a length of 15-40 amino acids.

[0242] In some examples, the peptides presented by the HLA protein comprise peptides identified by (a) isolating one or more HLA complexes from a cell line expressing a single HLA class II allele; (b) isolating one or more HLA-peptides from the one or more isolated HLA complexes; (c) obtaining MS/MS spectra for the one or more isolated HLA-peptides; and (d) obtaining a peptide sequence that corresponds to the MS/MS spectra of the one or more isolated HLA-peptides from a peptide database; wherein one or more sequences obtained from step (d) identifies the sequence of the one or more isolated HLA-peptides.

[0243] In some examples, the personalized cancer therapy further comprises an adjuvant.

[0244] In some examples, the personalized cancer therapy further comprises an immune checkpoint inhibitor.

[0245] In some examples, the training data comprises structured data, time-series data, unstructured data, relational data, or any combination thereof.

[0246] In some examples, the unstructured data comprises image data.

[0247] In some examples, the relational data comprises data from a customer system, an enterprise system, an operational system, a website, web accessible application program interface (API), or any combination thereof.

[0248] In some examples, the training data is uploaded to a cloud-based database.

[0249] In some examples, the training is performed using convolutional neural networks.

[0250] In some examples, the convolutional neural networks comprise at least two convolutional layers.

[0251] In some examples, the convolutional neural networks (CNN) comprise at least one batch normalization step.

[0252] In some examples, the convolutional neural networks comprise at least one spatial dropout step.

[0253] In some examples, the convolutional neural networks comprise at least one global max pooling step.

[0254] In some examples, the convolutional neural networks comprise at least one dense layer.

[0255] In some examples, identifying peptide sequences comprises identifying peptide sequences with a mutation expressed in cancer cells of a subject.

[0256] In some examples, identifying peptide sequences comprises identifying peptide sequences not expressed in normal cells of a subject.

[0257] In some examples, identifying peptide sequences comprises identifying overexpressed peptide sequences.

[0258] In some examples, identifying peptide sequences comprises identifying viral peptide sequences. In one example, provided herein is a method for identifying HLA class II specific peptides for immunotherapy specific for a subject, the method comprising: identifying a candidate peptide comprising an epitope; inputting amino acid information of a plurality of peptide sequences, each comprising an epitope, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of HLA presentation predictions for the peptide sequence to an immune cell, each presentation prediction representing a probability that one or more proteins encoded by an HLA class II allele of a cell of the subject will present a given peptide sequence comprising the epitope; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50%, selecting a protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject, predicted to bind to the candidate peptide by the prediction model, wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the candidate peptide to an immune cell; contacting the candidate peptide with the protein encoded by the HLA class II allele, such that the candidate peptide competes with a placeholder peptide associated with the protein encoded by the HLA class II allele; and, identifying the candidate peptide as a peptide for immunotherapy specific for the protein encoded by an HLA class II allele based on whether the candidate peptide displaces the placeholder peptide.

[0259] In some examples, the immunotherapy is cancer immunotherapy.

[0260] In some examples, identifying comprises comparing DNA, RNA or protein sequences from the cancer cells of the subject to DNA, RNA or protein sequences from the normal cells of the subject. In some embodiments, the epitope is a cancer specific epitope.

[0261] In some examples, the at least one protein encoded by the HLA class II allele comprises at least an alpha 1 subunit and a beta 1 subunit of the HLA protein, or fragments thereof, present in dimer form. In some examples, the placeholder peptide is a CLIP peptide. In some examples, the placeholder peptide is a CMV peptide. In some examples, the method further comprises measuring the IC50 of displacement of the placeholder peptide by the target peptide. In some examples, the IC50 of displacement of the placeholder peptide by the target peptide is less than 500 nM. In some examples, the at least one protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject is an HLA class II tetramer or multimer. In some examples, the target peptide is further identified by mass spectrometry. In some examples, the at least one protein encoded by the HLA class II allele of a cell of the subject is a recombinant

protein. In some examples, the at least one protein encoded by the HLA class II allele of a cell of the subject is expressed in a eukaryotic cell.

[0262] In one example, provided herein is assay method for verifying the specificity of a candidate peptide for binding an HLA class II protein, the method comprising: expressing in a eukaryotic cell, a polynucleic acid construct comprising a nucleic acid sequence encoding an HLA class II protein comprising an alpha chain and beta chain or portions thereof, capable of binding a peptide comprising an MHC-II-binding epitope, and wherein the expressed HLA class II protein or portions thereof remains associated with a placeholder peptide; isolating the HLA class II protein or portions thereof expressed in the eukaryotic cell; performing a peptide exchange assay by (a) adding increasing amount of the candidate peptide to determine whether the candidate peptide displaces the placeholder peptide associated with the HLA class II protein or portions thereof; and (b) calculating the IC50 of the displacement reaction to determine the affinity of the candidate peptide to the HLA class II protein or portions thereof relative to the placeholder peptide, thereby verifying the specificity of the candidate peptide for binding an HLA class II protein.

[0263] In some examples, the identity of the peptide is known. In some embodiments, the identity of the peptide is not known. In some examples, the identity of the peptide is determined by mass spectrometry.

[0264] In some examples, the peptide exchange assay comprises detection of peptide fluorescent probes or tags. In some examples, the placeholder peptide is a CLIP peptide.

[0265] In some examples, the polynucleic acid construct comprises an expression vector, further comprising one or more of: a promoter, a linker, one or more protease cleavage sites, a secretion signal, dimerization factors, ribosomal skipping sequence, one or more tags for purification and or detection.

[0266] In one example, provided herein is a method for assaying immunogenicity of a MHC class II binding peptide, the method comprising: selecting a protein encoded by an HLA class II allele predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50% and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence; contacting the peptide with the selected protein encoded by the HLA class II allele such that the peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele, and displaces the placeholder peptide, thereby forming a complex comprising the HLA class II protein and the identified peptide; contacting the HLA class II protein and the identified peptide complex with a CD4+ T cell, assaying for one or more of activation parameters of the CD4+T cell, selected from induction of a cytokine, induction of a chemokine and expression of a cell surface marker.

[0267] In some examples, the HLA class II allele is a tetramer or multimer. In some embodiments, the cytokine is IL-2. In some examples, the cytokine is IFN-gamma.

[0268] In one example, provided herein is a method for inducing a CD4+ T cells activation in a subject for cancer immunotherapy, the method comprising: identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying comprises comparing DNA, RNA or protein sequences from the cancer cells of the subject to DNA, RNA or protein sequences from the normal cells of the subject; selecting a protein

encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at the most 50%and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence; contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; purifying the identified peptide; and administer an effective amount of the identified peptide to the subject.

[0269] In one example provided herein is a method of manufacturing HLA class II tetramers or multimers, the method comprising: expressing in a eukaryotic cell, a vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a linker, one or more protease cleavage sites, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta 1 heterodimers is secreted in dimerized state, wherein the heterodimer is associated with a placeholder peptide, purifying the secreted heterodimer from cell medium, validating the peptide binding activity using peptide exchange assay, adding streptavidin thereby conjugating heterodimers into tetramers, purifying the tetramers and having an yield of greater than 1 mg/L.

[0270] In some examples, the vector comprises a CMV promoter. In some examples, the vector comprises a sequence encoding a placeholder peptide linked via a cleavable site to the beta1 chain. In some examples, peptide exchange assay involves prior cleavage of the placeholder peptide from the beta chain. In some examples, the cleavable site is a thrombin cleavage site. In some examples, peptide exchange assay is a FRET assay. In some examples, the purification is by any one of: column chromatography, batch chromatography, ion exchange chromatography, size exclusion chromatography, affinity chromatography or LC-MS.

[0271] In one example, provided herein is a composition comprising HLA class II tetramers comprising either HLA-DR, or HLA-DP, or HLA-DQ heterodimers, each heterodimer comprising an alpha and a beta chain, purified and present at a concentration of greater than 0.25 mg/L. In some examples, the HLA class II tetramer comprises heterodimer pairs selected from a group consisting of: protein may be selected from the group consisting of an HLA-DR, and HLA-DP or an HLA-DQ protein. In some examples, the HLA protein is selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01,HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02,HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-

DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, HLA-DRB5*01:01).

[0272] In some examples, the heterodimer pairs are expressed in a eukaryotic cell. In some examples, the heterodimer pair is encoded by a vector. In some examples, the vector comprises: a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, wherein the secreted heterodimer is associated with a placeholder peptide. In some examples, the vector comprises: a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, wherein the secreted heterodimer is associated with a placeholder peptide.

[0273] In some examples, HLA class II heterodimers secreted from eukaryotic cells into cell culture medium, and is purified by any one of: column or batch chromatography, ion exchange chromatography, size exclusion chromatography, affinity chromatography or LC-MS.

[0274] In one example, provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising: inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by an HLA class I or II allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data wherein the training data comprises: sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; (b) determining or predicting that each of the peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the set of presentation predictions; and (c) administering to the subject a composition comprising the drug.

[0275] In one example, provided herein is a method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising: (a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by an HLA class I or II allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-

peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data; wherein the training data comprises: sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; (b) determining or predicting that at least one of the peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the set of presentation predictions.

[0276] In one example, the method further comprises deciding not to administer the drug to the subject.

[0277] In one example, the drug comprises and antibody or binding fragment thereof.

[0278] In one example, the peptide sequences of the polypeptide sequences comprise each contiguous peptide sequence of the polypeptide sequence that has a length of 8, 9, 10, 11 or 12 amino acids, and wherein the protein encoded by an HLA class I or II allele of a cell of the subject is a protein encoded by an HLA class I allele of a cell of the subject.

[0279] In one example, the peptide sequences of the polypeptide sequences comprise each contiguous peptide sequence of the polypeptide sequence that has a length of 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 amino acids, and wherein the protein encoded by an HLA class I or II allele of a cell of the subject is a protein encoded by a class II MHC allele of a cell of the subject.

[0280] In one example, provided herein is a method of treating a subject with an autoimmune disease or condition comprising: (a) identifying or predicting an epitope of an expressed protein presented by an HLA class I or II of a cell of the subject, wherein a complex comprising the identified or predicted epitope and the HLA class I or II is targeted by a CD8 or CD4 T cell of the subject; (b) identifying a T cell receptor (TCR) that binds to the complex; (c) expressing the TCR in a regulatory T cell from the subject or an allogeneic regulatory T cell; and (d) administering the regulatory T cell expressing the TCR to the subject.

[0281] In one example, the autoimmune disease or condition is diabetes.

[0282] In one example, the cell is an islet cell.

[0283] In one example, provided herein is a method of treating a subject with an autoimmune disease or condition comprising administering to the subject a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising (i) an epitope of an expressed protein identified or predicted to be presented by an HLA class I or II of a cell of the subject and (ii) the HLA class I or II, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.

[0284] Additional examples and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative examples of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different examples, and its several details are capable of modifications in various obvious respects, all without

departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

[0285] MAPTAC™ can be used for high-throughput peptide binding assays where peptides bound to HLA class II are measured after isolation with MAPTAC™ constructs at different time points and under different conditions, such as heating at 37°C, to obtain the sequences of populations of peptides with different stabilities using LC-MS/MS.

[0286] In one example, provided herein is a method for treating a cancer in a subject the method comprising: identifying peptide sequences, wherein the peptide sequences are associated with cancer, wherein identifying comprises comparing DNA, RNA or protein sequences from the cancer cells of the subject to DNA, RNA or protein sequences from the normal cells of the subject; inputting amino acid information of the peptide sequences identified, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences identified, each presentation prediction representing a probability that one or more proteins encoded by an HLA class II allele of a cell of the subject will present a given sequence of a peptide sequence identified; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data wherein the training data comprises: sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and selecting a subset of the peptide sequences identified based on the set of presentation predictions for preparing the personalized cancer therapy; and administering to the subject a composition comprising one or more of the peptides, wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at most 50%.

[0287] In some examples, the machine-learning HLA-peptide presentation prediction model comprises sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry after performing reverse phase offline fractionation.

[0288] In some examples, the prediction model exhibits a 1.1x to 100x fold improvement compared to NetMHCIIpan. In some examples, the prediction model exhibits a 1.1, 2, 3, 4, 5, 6, 7, 7.4, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 50, 55, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 8, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 -fold or more improvement compared to NetMHCIIpan.

INCORPORATION BY REFERENCE

[0289] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference. To the extent publications and patents or patent

applications incorporated by reference contradict the disclosure contained in the specification, the specification is intended to supersede and/or take precedence over any such contradictory material.

BRIEF DESCRIPTION OF THE DRAWINGS

[0290] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "FIG." herein), of which:

[0291] FIG. 1A diagram representing a peptide docked onto MHC Class I protein. Figure discloses SEQ ID NO: 36.

[0292] FIG. 1B depicts an exemplary diagram representing a peptide docked onto MHC Class II protein. Figure discloses SEQ ID NO: 37.

[0293] FIG. 2 depicts an exemplary experimental approach for generating mono-allelic HLA class II binding peptide data. HLA class II peptides are introduced into any cell, including a cell not expressing HLA class II so that specific HLA class II allele(s) are expressed in the cell. Populations of genetically engineered HLA expressing cells are harvested, lysed, and their HLA-peptide complexes are tagged (*e.g.*, biotinylated) and immunopurified (*e.g.*, using the biotin-streptavidin interaction). HLA-associated peptides specific to a single HLA can be eluted from their tagged (*e.g.*, biotinylated) complexes and evaluated (*e.g.*, sequenced using high resolution LC-MS/MS).

[0294] FIG. 3 depicts an exemplary sequence logo representation of HLA class II-DRB1*11:01 -associated peptides across Neon BAP, Expi293 cell line; Neon BAP, A375 cell line; IEDB, Affinity < 50 nM; and Pan-HLA Class II Ab, Homozygous LCL. **FIG. 3** shows that examples of MS-derived motifs match known patterns and show consistency across transfected cell lines.

[0295] FIG. 4 is an exemplary depiction of the HLA class II binding predictor performance. **FIG. 4** is a bar plot showing the performance of the binding predictor (*neonmhc2*) and NetMHCIIpan applied to a validation dataset consisting of observed mass spec peptides and decoy peptides which are generated at a ratio of 1:19 (hits:decoys) by randomly shuffling the hit peptides. For the NEON binding predictor *neonmhc2*, a separate model is built for each MHC II allele shown. The height of the bars shows the positive predictive value (PPV), defined as the fraction of predicted binders in the validation set which were indeed hit peptides. The alleles are sorted by the model's performance when predicting for that allele.

[0296] FIG. 5 depicts an exemplary effect of scored peak intensity (SPI) thresholds on binding predictor validation. FIG. 5 shows the performance of the HLA class II binding predictor when trained/validated on sets of peptides with different scored peak intensity (SPI) cutoffs. For each allele-specific model that is trained, shown is the model's performance in 3 settings: trained and evaluated on datasets using observed MS hit peptides of larger than or equal to 70 SPI, trained on peptides with larger than or equal to 50 SPI and validated on peptides with larger than or equal to 70 SPI, and trained and validated on peptides with larger than or equal to 50 SPI.

[0297] FIG. 6 depicts an exemplary bar plot showing *representative* data from number of observed peptides by allele profiling by LC-MS/MS with larger than or equal to 70 scored peak intensity (SPI) cutoffs. Each bar represents the total number of observed peptides of an allele. There are collected data for 35 HLA-DR alleles.

The collected data for 35 HLA-DR alleles have >95% population coverage for HLA-DR (USA allele frequencies).

[0298] FIG. 7A shows the PPV of the model when applied to test partition of data for the indicated HLA class II alleles. The decoy peptides used were scrambled sequences of the positive (hit) peptide sequences at a hit to decoy ratio of 1:19. PPV was determined by identifying the top-scoring 5% of peptides in the test partition and determining the fraction of them that were positive for binding to the protein encoded by the respective HLA class II allele.

[0299] FIG. 7B depicts exemplary prediction performance as a function of training set size (curves obtained by artificially down-sampling the training set). FIG. 7B shows that, generally, for the 35 HLA-DR alleles collected, when the training set size increases, the value of PPV increases.

[0300] FIG. 8 depicts an exemplary graph, demonstrating that processing-related variables can improve prediction further. Distinguish MS-observed peptides random sequences selected from protein-coding exome may be distinguished. On the training data partition, a logistic regression may be fit to predict HLA class II presentation using binding strength (NetMHCIIpan or Neon's predictor) and processing features (RNA-Seq expression and a derived gene-level bias term). On a separate evaluation partition, exonic positions overlapping MS-observed MHC II peptides ("hits") may be scored alongside random exonic positions not observed in MS (1:499 ratio). The top 0.2% (1/500) may be called as positives, and positive predictive value may be assessed this threshold.

[0301] FIG. 9 depicts an exemplary neural network architecture. Input peptides are represented as 20mers, with shorter peptides being filled in with "missing" characters. Each peptide has a 31-dimensional embedding, so the input into the neural network is a 20x31 matrix. Before being processed by the neural network, feature normalization on the 20x31 matrix is performed based on feature value means and standard deviations in the training set. The first convolutional layer has a kernel of 9 amino acids and 50 filters (also called channels) with a Rectified Linear Unit (ReLU) activation function. This is followed by batch normalization then spatial dropout with a dropout rate of 20%. This is followed by another convolutional layer with a kernel of 3 amino acids and 20 filters with a ReLU activation function and then again followed by batch normalization and spatial dropout with a dropout rate of 20%. Global max pooling is then applied, taking the maximally-activated neuron in each of the 20 filters; then these 20 values are passed into a fully connected (dense) layer with a single neuron using a Sigmoid activation function. The output of this layer is treated as the binding/non-binding prediction. L2 regularization is applied to the weights of the first convolutional layer, second convolutional layer, and dense layer with weights of 0.05, 0.1, and 0.01, respectively. Additional models used have varied the number of convolutional layers and the kernel size of each layer.

[0302] FIG. 10 depicts an exemplary computer control system that is programmed or otherwise configured to implement methods provided herein.

[0303] FIG. 11A depicts an exemplary overview of the MAPTAC™ experimental workflow. Figure discloses SEQ ID NO: 38.

[0304] FIG. 11B depicts exemplary per-allele peptide counts, merged across replicates.

[0337] FIG. 21D depicts an exemplary diagram representing overlap of heavy-labeled peptide source genes according to the lysate and UV-treatment experiments. Gene names are colored by functional class.

[0338] FIG. 22A depicts an exemplary flow diagram representing an assay protocol disclosed herein, to validate HLA class II- driven CD4+T cells and T cell responses.

[0339] FIG. **22B** depicts an exemplary HLA protein dimer construct design for peptide exchange assay (upper panel) and a graphical representation of an exemplary assay workflow (lower panel). Figure discloses "10XHis" as SEQ ID NO: 20.

[0340] FIG. 23 depicts an exemplary graphical illustration of an exemplary vector design for MHC-II expression for screening new binding peptides, and a representation of the expressed protein product. Figure discloses SEQ ID NO: 39 and discloses "10XHis" as SEQ ID NO: 20.

[0341] FIG. 24 depicts an exemplary flow diagram of transfection, purification and cleavage of placeholder peptide from beta chain.

[0342] FIG. 25A depicts an exemplary graphical illustration showing vector encoding CLIP peptides that are associated with increased secretion of expressed MHC-II peptides. Figure discloses SEQ ID NO: 21.

[0343] FIG. 25B depicts an exemplary graphical representation with the shorter and longer forms of the nucleic acids encoding CLIP0 and CLIP1 respectively. Figure discloses SEQ ID NOS 1 and 21, respectively, in order of appearance.

[0344] FIG. 25C depicts an exemplary representative result of a Coomassie gel analysis of the alpha and beta chains with or without the longer clip.

[0345] FIG. 26A depicts an exemplary graphical illustration of the TR-FRET assay.

[0346] FIG. 26B depicts exemplary representative polarization data from an HLA class II peptide binding assay using Fluorescence Resonance Energy Transfer (FRET) assay using specific peptides.

[0347] FIG. 26C depicts exemplary representative polarization data from an HLA class II peptide binding assay using Fluorescence Resonance Energy Transfer (FRET) assay using specific peptides.

[0348] FIG. 26D depicts an exemplary percent displacement of MHC-construct bound peptide that was calculated from increase in fluorescence**.**

[0349] FIG. 26E depicts an exemplary percent displacement of MHC-construct bound peptide that was calculated from increase in fluorescence.

[0350] FIG. 26F depicts an exemplary peptide exchange using assay using differential scanning fluorometry (DSF). A graphical representation is depicted showing an exemplary mechanism of detecting peptide dissociation from MHC class II with heat which also dissociates the MHC class II heterodimer, resulting in binding of the fluorophore and high fluorescence. An exemplary schematic of placeholder peptide dislodgement by epitope peptide is also depicted. Exemplary melting curves plotted over temperature are also depicted.

[0351] FIG. **26G** depicts an exemplary soluble HLA-DM construct and its use for the performance of MHC Class II peptide exchange. The construct depicted contains a CMV promoter, a coding sequence for HLA-DM beta chain and a coding sequence for a HLA-DM alpha chain downstream of a secretion sequence (leader) and a BAP sequence at the 3'end of the beta chain coding sequence; a His tag at the 3'end of the alpha chain coding sequence. The two chains are be separated by an intervening ribosomal skipping sequence. The construct was expressed in Expi-CHO cells and the protein secreted into the medium culture medium was purified. Figure discloses "10XHis" as SEQ ID NO: 20.

[0366] FIG. 29 depicts an exemplary average count of peptides from an average MAPTAC™ experimental replicate (50 million cells), per each HLA allele.

[0367] FIGs. 30A-30C depict an exemplary binding core analysis for HLA class II MAPTAC™ alleles +/- HLA-DM and multi-allelic deconvolution fidelity. **FIG. 30A** depicts exemplary sequence logos for one representative HLA-DR, -DQ, and -DP allele according to MAPTAC™ with and without HLA-DM co-transfection (expi293 cell line) and IEDB wherein the height of each amino acid is proportional to its frequency. Amino acids with frequency greater than 10% are shown in color according to chemical properties; all others are shown in gray. Peptides were aligned according to the GibbsCluster tool (**Supplemental Methods**), and logos represent all peptides, including those that did not closely match the overall motif (e.g. no peptides are sequestered in a "trash" cluster). **FIG. 30B** depicts an exemplary description of cluster assignments for MAPTAC™ peptides (20 per allele) spiked into pan-DR MS datasets. Datasets were deconvolved using GibbsCluster. Each colored box represents one MAPTAC™ peptide. The color of the box indicates which cluster it was assigned to, and gray bars indicate which allele the peptide came from. **FIG. 30C** depicts an exemplary graph showing that the share of peptides exhibiting 0, 1, 2, 3, or 4 expected residues in anchor positions, for alleles shown in **FIG. 30B**. Anchor positions were defined as the four positions with lowest entropy, and the "expected" residues were defined as those with ≥10% frequency in those positions.

[0368] FIGs. 31A-31F depict an exemplary architecture and benchmarking of the *neonmhc2* binding prediction algorithm**. FIG. 31A** depicts an exemplary architecture of a convolutional neural network (CNN) trained to distinguish mono-allelic HLA class II peptides from scrambled length-matched decoys. The schematic indicates the usage of an amino acid feature embedding layer, 2 convolutional layers of width 6, the presence of skip-to-end connections, and a combination of average- and max-pooling operations as input to a final logistic output node. **FIG. 31B** depicts an exemplary positive predictive value (PPV) for NetMHCIIpan and *neonmhc2* as evaluated on a partition of MAPTAC™ data that was not used for training or hyper-parameter optimization. For each allele, *n* MS-observed peptides were scored in conjunction with 19*n* length-matched decoys sampled from the same set of source genes, and each predictor's *n* top-ranked peptides (e.g. the top 5%) were called as positives. According to this evaluation protocol, PPV is identical to recall because the number of false positives and false negatives is necessarily equal. **FIG. 31C** depicts an exemplary PPV for NetMHCIIpan and *neonmhc2* on the TGEM data set. For each allele, the *n* top-ranked peptides were called positives, where *n* is the number of confirmed immunogenic epitopes in the evaluated set. **FIG. 31D** depicts exemplary *ex vivo* T cell induction results for neoantigen peptides. Peptides were selected based on high *neonmhc2* scores and weak NetMHCIIpan scores for HLA-DRB1*11:01. Figure discloses SEQ ID NOS 87-89, 91, 90, 2, 92-94, 3, and 95-96, respectively, in order of appearance. **FIG. 31E** depicts comparison of models trained on monoallelic MAPTAC data versus deconvolved multiallelic data as evaluated on hold-out monoallelic data. Values are as shown for *neonmhc2* where the training dataset is down-sampled to match the size of the deconvolution training set. **FIG. 31F** shows PPV on the TGEM dataset for NetMHCIIpan-v3.1, the deconvolution-trained predictor, and *neonmhc2* (with and without down-sampling). For each allele, the *n* top -ranked peptides were called positives, where *n* is the number of confirmed immunogenic epitopes in the evaluated set.

depicts significance of various gene localizations and functional classes in predicting heavy (K562-derived and light (DC-derived) peptides respectively. P-values are calculated according to logistic regression that controls for *neonmhc2* binding score and source gene expression. Bar colors indicate sign associated with coefficient in the regression. **FIG. 33F** depicts an exemplary graphical representation of results showing overlap of tumor cell-derived peptide source genes (colored by functional class) in the UV- and HOCl-treated experiments. **FIG. 33G** depicts exemplary data showing PPV for predicting presented tumor antigens in a second donor using logistic models fit on heavy-labeled peptides observed in the first donor. Models were fit using *neonmhc2* binding alone; binding and expression; or binding, expression, and a binary variable indicating if a peptide was from a mitochondria gene.

[0371] FIGs. 34A-34B depict exemplary characterization of MAPTAC™ data related to **FIG. 29**. **FIG. 34A** depicts an exemplary HLA cell surface analysis by FACS of Expi293 cell lines transfected with MAPTAC™ constructs coding for affinity-tagged HLA-A*02:01-BAP **FIG. 34B** depicts an exemplary HLA cell surface analysis by FACS of Expi293 cell lines transfected with MAPTAC™ constructs coding for affinity-tagged HLA-DRB1*11:01-BAP (bottom). HLA cell surface expression of transfected Expi293 cells (orange) were compared with stained untransfected Expi293 (blue), unstained untransfected Expi293 (red), stained PBMCs (dark green), and unstained PBMCs (light green). All HLA class I stains utilized W6/32 (pan-HLA class I), while HLA class II stains utilized REA332 (pan-HLA class II).

[0372] FIG. 35 depicts an exemplary comparison of MAPTAC™ and IEDB logos, related to **FIG. 30A**. Measured and NetMHCIIpan-predicted affinities for MS-observed peptides that did not exhibit good NetMHCIIpan scores but were well supported by MS (scored peak intensity >70 and nested set size ≥1).

[0373] FIGs. 36A-36C depict an exemplary analysis of HLA-DR1 MAPTAC™ data fidelity, related to FIG. 30A-30C. **FIG. 36A** depicts exemplary NetMHCIIpan3.1 scores for HLA-DR1 MAPTAC™ peptides(green) (lengths 12-23) as compared to 50,000 length-matched decoy peptides randomly sampled from the proteome(blue), for common alleles. **FIG. 36B** depicts exemplary measured and NetMHCIIpan-predicted affinities for exemplary MS-observed peptides that did not exhibit good NetMHCIIpan scores but were well-supported by MS (scored peak intensity>70 and nested set size $\underline{\geq}$1). Figure discloses SEQ ID NOS 40-86, top to bottom, left to right, respectively, in order of appearance. **FIG. 36C** depicts exemplary HLA class II sequence logos for HLA-DRB1 alleles as determined by MAPTAC™ in different cell types.

[0374] FIGs. 37A-37C depict an additional exemplary analysis of MAPTAC™ motifs, related to **FIGs. 30A-30C**. **FIG. 37A** depicts MAPTAC™ -derived sequence logos for experiments with and without HLA-DM co-transfection (expi293 cell line). **FIG. 37B** depicts sequence logos for several HLA class I alleles according to MAPTAC™ and IEDB. Note that A*32:01 does not show a high frequency Q at P2 and C*03:03 does not show a high frequency Y at P9, differing with previous studies that used multi-allelic deconvolution; the logo for B*52:01 is previously unpublished. **FIG. 37C** depicts an exemplary alignment of MAPAC™-observed peptides to the gene sequence of CD74.

[0375] FIGs. 38Ai-38D depict exemplary *neonmhc2* performance statistics and T cell flow staining, related to **FIGs. 31A-31D**. **FIG. 38Ai** depicts an exemplary performance of *neonmhc2* as a function of training data set size. PPV was evaluated in the same manner and using the same evaluation peptides as in **FIG. 31B**; however,

the training data was randomly down-sampled to mimic smaller training data sets. **FIG. 38Aii** depicts exemplary sequence logos for peptide clusters derived from multi-allelic HLA-DR ligandome using GibbsCluster (default settings; "trash cluster allowed). **FIG. 38B** depicts exemplary representative flow cytometry plots of IFN-γ expression by CD4+ cells from induction samples recalled with neoantigen peptides predicted with *neonmhc2*. Delta values were calculated by subtracting the percent of CD4+ cells expressing IFN-γ when recalled with neoantigen (+Peptide) from the percent of CD4+ expressing IFN-γ when recalled in the presence of no neoantigen (No Peptide). The left two flow plots are representative of a neoantigen that induced a CD4+ T cell T cell response (PEASLYGALSKGSGG (SEQ ID NO: 2)) and a neoantigen that did not induce a T cell response (PATYILILKEFCLVG (SEQ ID NO: 3)). **FIG. 38C** depicts exemplary delta values from wells recalled with single *neonmch2* neoantigen peptides. Peptides were considered an induction hit if they had a positive response (delta response above 3%, highlighted). Figure discloses SEQ ID NOS 87-91, 2, 92-94, 3, and 95-96, respectively, in order of appearance. **FIG. 38D** shows exemplary sequence logos for peptide clusters derived for multi-allelic HLA-DR ligandomes using GibbsCluster (default settings; "trash" cluster allowed).

[0376] FIGs. 39A-39C depict an additional exemplary cell-of-origin analysis for HLA class II, related to **FIGs. 32A-32E**. **FIG. 39A** depicts exemplary percent-rank *neonmhc2* scores for HLA class II peptides observed in 4 PBMC samples profiled by pan-DR antibody (RG1248, RG1104, RG1095, and HDSC from **FIG. 30B**), according to whether the peptide source gene is present in human plasma. For each peptide, the best (lowest) percent rank was used across the alleles present in the donor. Scores for random length-matched proteome decoys are shown for comparison. Box plots mark the 5th, 25th, 50th, 75th, and 95th percentiles. **FIG. 39B** depicts exemplary counts of observed vs. expected peptides per gene for HLA class I, using the same methodology as in **FIG. 32A**. Data correspond to the same tumor types (colorectal, ovarian, and melanoma). Genes present in human plasma are highlighted in blue and sized according to their concentration. **FIG. 39C** depicts an exemplary relative concordance of peptide observations with respect to two different gene expression profiles. For each sample, gene-level peptide counts were modeled as a linear combination of a bulk tumor gene expression and professional APC gene expression profile. The ratio of the coefficients determines the relative concordance of each expression profile with the peptide repertoire. Error bars correspond to a 95% confidence interval computed by bootstrap resampling.

[0377] FIGs. 40A-40B depict an additional exemplary analysis of processing motifs related to **FIGs. 32A-32E**. **FIG. 40A** depicts exemplary amino acid frequencies near N-terminal and C-terminal peptide cut sites relative to average proteome frequencies (applies for upstream positions U3-U1 and downstream positions D1-D3) or relative to average peptide frequencies (applies for internal positions N1-C1) as observed in donor PBMC, monocyte-derived dendritic cells, colorectal cancer, melanoma, ovarian cancer, and the expi293 cell line (used for most MAPTAC™ data generation). **FIG. 40B** depicts the same analysis as **FIG. 32E** but using NetMHCIIpan as the base predictor. Absolute increase in PPV observed for logistic regression models that included processing-related variables in addition to NetMHCIIpan predictions (as compared to NetMHCIIpan-only models) for eight samples profiled by HLA-DR antibody (the same samples analyzed in **FIG. 31B**). Asterisks mark significant improvements (*: $p<0.01$, **: $p<0.001$, ***: $p<0.0001$) according to two-tailed paired t-tests.

required in order to achieve the desired therapeutic effect and gradually increase the dosage until the desired effect is achieved. Disease, condition, and disorder are used interchangeably herein.

[0426] Those of ordinary skill in the art will recognize that the terms “peptide tag,” “affinity tag,” “epitope tag,” or “affinity acceptor tag” are used interchangeably herein. As used herein, the term “affinity acceptor tag” refers to an amino acid sequence that permits the tagged protein to be readily detected or purified, for example, by affinity purification. An affinity acceptor tag is generally (but need not be) placed at or near the N- or C-terminus of an HLA allele. Various peptide tags are well known in the art. Non-limiting examples include poly-histidine tag (e.g., 4 to 15 consecutive His residues (SEQ ID NO: 4), such as 8 consecutive His residues (SEQ ID NO: 5)); poly-histidine-glycine tag; HA tag (e.g., Field et al., Mol. Cell. Biol., 8:2159, 1988); c-myc tag (e.g., Evans et al., Mol. Cell. Biol., 5:3610, 1985); Herpes simplex virus glycoprotein D (gD) tag (e.g., Paborsky et al., Protein Engineering, 3:547, 1990); FLAG tag (e.g., Hopp et al., BioTechnology, 6:1204, 1988; U.S. Pat. Nos. 4,703,004 and 4,851,341); KT3 epitope tag (e.g., Martine et al., Science, 255:192, 1992); tubulin epitope tag (e.g., Skinner, Biol. Chem., 266:15173, 1991); T7 gene 10 protein peptide tag (e.g., Lutz-Freyemuth et al., Proc. Natl. Acad. Sci. USA, 87:6393, 1990); streptavidin tag (StrepTag.TM. or StrepTagII.TM.; see, e.g., Schmidt et al., J. Mol. Biol., 255(5):753-766, 1996 or U.S. Pat. No. 5,506,121; also commercially available from Sigma-Genosys); or a VSV-G epitope tag derived from the Vesicular Stomatis viral glycoprotein; or a V5 tag derived from a small epitope (Pk) found on the P and V proteins of the paramyxovirus of simian virus 5 (SV5). In some embodiments, the affinity acceptor tag is an “epitope tag,” which is a type of peptide tag that adds a recognizable epitope (antibody binding site) to the HLA-protein to provide binding of corresponding antibody, thereby allowing identification or affinity purification of the tagged protein. Non-limiting example of an epitope tag is protein A or protein G, which binds to IgG. In some embodiments, the matrix of IgG Sepharose 6 Fast Flow chromatography resin is covalently coupled to human IgG. This resin allows high flow rates, for rapid and convenient purification of a protein tagged with protein A. Numerous other tag moieties are known to, and can be envisioned by, the ordinarily skilled artisan, and are contemplated herein. Any peptide tag can be used as long as it is capable of being expressed as an element of an affinity acceptor tagged HLA-peptide complex.

[0427] As used herein, the term “affinity molecule” refers to a molecule or a ligand that binds with chemical specificity to an affinity acceptor peptide. Chemical specificity is the ability of a protein's binding site to bind specific ligands. The fewer ligands a protein can bind, the greater its specificity. Specificity describes the strength of binding between a given protein and ligand. This relationship can be described by a dissociation constant (KD), which characterizes the balance between bound and unbound states for the protein-ligand system.

[0428] The term “affinity acceptor tagged HLA-peptide complex” refers to a complex comprising an HLA class I or class II-associated peptide or a portion thereof specifically bound to a single allelic recombinant HLA class I or class II peptide comprising an affinity acceptor peptide.

[0429] The terms “specific binding” or “specifically binding” when used in reference to the interaction of an affinity molecule and an affinity acceptor tag or an epitope and an HLA peptide mean that the interaction is dependent upon the presence of a particular structure (e.g., the antigenic determinant or epitope) on the protein;

embodiments, the cleavage sequence is a T2A, P2A, E2A, or F2A sequence. For example, the cleavage sequence can be E G R G S L T C G D V E N P G P (SEQ ID NO: 6) (T2A), A T N F S L K Q A G D V E N P G P (SEQ ID NO: 7) (P2A), Q C T N Y A L K L A G D V E S N P G P (SEQ ID NO: 8) (E2A), or V K Q T L N F D L K L A G D V E S N P G P (SEQ ID NO: 9) (F2A).

[0535] In some embodiments, the cleavage sequence may be a thrombin cleavage site CLIP.

[0536] The peptides presented by the HLA protein may comprise peptides that are identified by searching a no-enzyme specificity without modification peptide database. The peptide database may be a no-enzyme specificity peptide database, such as a without modification database or a with modification (*e.g.*, phosphorylation or cysteinylation) database. In some embodiments, the peptide database is a polypeptide database. In some embodiments, the polypeptide database may be a protein database. In some embodiments, the method further comprises searching the peptide database using a reversed-database search strategy. In some embodiments, the method further comprises searching a protein database using a reversed-database search strategy. In some embodiments, a *de novo* search is performed, *e.g.*, to discover new peptides that are not included in a normal peptide or protein database. The peptide database may be generated by providing a first and a second population of cells each comprising one or more cells comprising an affinity acceptor tagged HLA, wherein the sequence affinity acceptor tagged HLA comprises a different recombinant polypeptide encoded by a different HLA allele operatively linked to an affinity acceptor peptide; enriching for affinity acceptor tagged HLA-peptide complexes; characterizing a peptide or a portion thereof bound to an affinity acceptor tagged HLA-peptide complex from the enriching; and generating an HLA-allele specific peptide database.

[0537] The peptides presented by the HLA protein may comprise peptides identified by comparing a MS/MS spectra of the HLA-peptides with MS/MS spectra of one or more HLA-peptides in a peptide database.

[0538] There may be mutation on either peptides or nucleic acid that encodes peptides. The mutation may be selected from the group consisting of a point mutation, a splice site mutation, a frameshift mutation, a read-through mutation, and a gene fusion mutation. The point mutation may be a genetic mutation where a single nucleotide base is changed, inserted or deleted from a sequence of DNA or RNA. The splice site mutation may be a genetic mutation that inserts, deletes or changes a number of nucleotides in the specific site at which splicing takes place during the processing of precursor messenger RNA into mature messenger RNA. The frameshift mutation may be a genetic mutation caused by indels (insertions or deletions) of a number of nucleotides in a DNA sequence that is not divisible by three. The mutation may also comprise insertions, deletions, substitution mutations, gene duplications, chromosomal translocations, and chromosomal inversions.

[0539] In some embodiments, the HLA class II protein comprises an HLA-DR protein.

[0540] In some embodiments, the HLA class II protein comprises an HLA-DP protein.

[0541] In some embodiments, the HLA class II protein comprises an HLA-DQ protein.

[0542] In some embodiments, the HLA class II protein may be selected from the group consisting an HLA-DR, and HLA-DP or an HLA-DQ protein. In some embodiments, the HLA protein is an HLA class II protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-

[0656] With regards to **FIG. 12A** and **FIG. 12E**, the mhc_ligand_full.csv dataset was downloaded from IEDB data on September 21, 2018. Valid affinity measurements were required to have a "Method/Technique" equal to "cellular MHC/competitive/fluorescence", "cellular MHC/competitive/radioactivity", "cellular MHC/direct/fluorescence", "purified MHC/competitive/fluorescence", "purified MHC/competitive/radioactivity", or "purified MHC/direct/fluorescence" and an "Assay Group" equal to "dissociation constant KD", "dissociation constant KD (~EC50)", "dissociation constant KD (~IC50)", "half maximal effective concentration (EC50)", or "half maximal inhibitory concentration (IC50)". A measurement was attributed to the Søren Buus group (University of Copenhagen, Denmark) if the string "Buus" appeared in the "Authors" field. Otherwise, if the authors field included the strings "Sette" or "Sidney", a measurement was attributed to the Alessandro Sette group (La Jolla Institute for Immunology, U.S.A). All other measurements were labeled as "Other". For the purposes of enumerating strong binders, only peptides with a measured affinity stronger than 50 nM were counted (**FIG. 12A**). **FIG. 12E** includes additional data from tools.iedb.org/main/datasets/, and strong binders with affinity <100 nM are enumerated.

DNA construct design

[0657] The gene sequences for *HLA* class I and *HLA* class II alleles were identified by the IPD-IMGT/HLA webpage (ebi.ac.uk/ipd/imgt/hla) and used to design recombinant expression constructs. For *HLA* class I, the α-chain was fused with a C-terminal GSGGSGGSAGG linker (SEQ ID NO: 10), followed by the biotin-acceptor-peptide (BAP) tag sequence GLNDIFEAQKIEWHE (SEQ ID NO: 11), a stop codon, and a variable DNA barcode, and cloned into the pSF Lenti vector (Oxford Genetics, Oxford, UK) via the NcoI and XbaI restriction sites. The *HLA* class II constructs were similarly cloned into pSF Lenti via the NcoI and XbaI restriction sites and consisted of the β-chain sequence fused on the C-terminus to the linker-BAP sequence from the class I construct (SGGSGGSAGGGLNDIFEAQKIEWHE (SEQ ID NO: 12)), followed by another short GSG linker an a F2A ribosomal skipping sequence (VKQTLNFDLLKLAGDVESNPGP (SEQ ID NO: 13)), the sequence of the α-chain, an HA tag (GSYPYDVPDYA (SEQ ID NO: 14)), a stop codon, and a variable DNA barcode. The identity of all DNA sequences was verified by Sanger sequencing.

Cell culture and transient transfections

[0658] Expi293 cells (Thermo Scientific) were grown in Expi293 medium (Thermo Scientific) with 8% CO_2 at 37°C with shaking at 125 rpm. Expi293 cells were maintained at cell densities between 0.5 x 10^6/mL and 6 x 10^6/mL with regular biweekly passaging. 30 mL of the Expi293 cell suspension was used for transient transfections at a cell density of approximately 3x10^6/mL and >90% viability. Briefly, 30 ug DNA (1 µg/mL DNA per mL cell suspension) was diluted into 1.5 mL Opti-MEM medium (Thermo Scientific) in one tube while 80 µL ExpiFectamine™ 293 transfection reagent (Thermo Scientific) was diluted into a second tube containing 1.5 mL Opti-MEM. These two tubes were incubated at room temperature for five minutes, combined, mixed gently, and incubated at room temperature for 30 minutes. The DNA and ExpiFectamine mixture was added to Expi293 cells and incubated at 37°C, 8% CO_2, 80% relative humidity. After 48 h, transfected cells were harvested in four technical replicates at 50x10^6 cells per tube, centrifuged, washed once with 1 x Gibco DPBS (Thermo Scientific), and flash frozen in liquid nitrogen for mass spectrometric analysis. An aliquot of 1 x 10^6 cells was collected from each transfection batch and analyzed via anti-BAP (Rockland Immunochemicals

Inc., Limerick, PA) or anti-HA (Bio-Rad, Hercules, CA) western blot to verify affinity-tagged HLA protein expression.

[0659] A375 cells (ATCC) were grown in DMEM with 10% FBS and maintained at cultures at no greater than 80% confluence with regular passaging. For mass spectrometry experiments A375 cells were cultured in a 500 cm^2 plate at a seeding density of 18.5 x 10^6 cells/mL in 100 mL, as calculated from a 70% confluent cell number. After 24 hours, cells were transfected with TransIT-X2 (Mirus Bio, Madison, WI) by following the TransIT system protocol adjusted for the total culture volume. After 48 h, cell medium was aspirated, and cells were washed with 1X Gibco DPBS (Thermo Scientific). For harvest, A375 cells were incubated for 10 minutes at 37°C with 30 mL non-enzymatic cell dissociation solution (Sigma-Aldrich), centrifuged, washed with 1X DPBS, and aliquoted at 50 x10^6 cells per sample. 293T and HeLa cells were purchased from ATCC and were cultured at 37°C at 5% CO2 in DMEM, 10% FBS, 2 mM L-glutamine or DMEM + 10% FBS, respectively. Both cell lines were transfected with the HLA constructs using the TransIT LT1 reagent (Mirus Bio, Madison, WI) following the manufactures instructions and processed 48h after transfection as described for the A375 cells. From all samples, an aliquot of 1 x 10^6 cells was collected from each transfection and analyzed via anti-BAP (Rockland Immunochemicals Inc., Limerick, PA) or anti-HA (Bio-Rad, Hercules, CA) western blot to verify affinity-tagged HLA protein expression.

BirA protein expression and purification

[0660] The pET19 vector encoding *E. coli* BirA fused to a C-terminal hexa-histidine tag (SEQ ID NO: 15) was used. Chemical competent *E. coli* BL21 (DE3) cells (New England Biolabs) were transformed with the BirA expression plasmid, grown at 37°C in LB broth plus 100 μg/ml ampicillin to an OD_{600} of 0.6–0.8 and cooled to 30°C before expression was induced by adding 0.4 mM isopropyl-β-D-thiogalactopyranoside. *E. coli* cell growth continued at 30°C for 4 h. *E. coli* cells were harvested by centrifugation at 8000 x g for 30 minutes at 4 °C and stored at –80°C until use. Frozen cell pellets expressing recombinant BirA were resuspended in IMAC buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl) with 5 mM Imidazole, incubated with 1 mg/ml lysozyme for 20 minutes on ice and the lysed by sonication. Cellular debris and insoluble materials were removed by centrifugation at 16,000 x g for 30 minutes at 4°C. The cleared supernatant was subsequently loaded on a HisTrap HP 5 mL column using the ÄKTA pure chromatography system (GE Healthcare), washed with IMAC buffer plus 25 mM and 50 mM imidazole before elution with 500 mM imidazole. Fractions containing BirA were pooled and dialyzed against 20 mM Tris-HCl pH 8.0 with 25 mM NaCl and were loaded on a HiTrap Q HP 5mL column (GE Healthcare) and eluted by applying a linear gradient from 25 to 600 mM NaCl. Fractions containing highly pure BirA were pooled, buffer exchanged in storage buffer (20 mM Tris-HCl pH 8.0 100 mM NaCl, 5% glycerol) and concentrated to around 5-10 mg/mL, aliquoted, and flash frozen in liquid nitrogen for storage at –80°C. BirA protein concentration was determined by UV spectroscopy at $OD_{280\ nm}$ using a calculated extinction coefficient of $\varepsilon = 47{,}440\ M^{-1}\ cm^{-1}$.

Western blotting protocol

variable range precursor mass filtering, and score and delta Rank1 – Rank2 score thresholds optimized across all LC-MS/MS runs for an HLA allele yielding a PSM FDR estimate of <1.0% for each precursor charge state.

[0670] Identified peptides that passed the PSM FDR estimate of <1.0% were further filtered for contaminants by removing all peptides assigned to the 264 common contaminants proteins in the reference database and by removing peptides identified in the negative control MAPTAC™ affinity pulldowns. Additionally, all peptide identifications that mapped to an *in silico* tryptic digest of the reference database were removed, as these peptides cannot be ruled out as tryptic contaminants from sample carry-over on the uPLC column.

Monoallelic Assignment of HLA-DR, -DQ, DP heterodimers using MAPTAC™ **protocol**

[0671] Mono-allelic HLA assignment to LC-MS/MS identified peptides followed two approaches. Because allelic variation in HLA-DRA1 is limited and not considered to influence peptide binding, all data from DR experiments (profiling DRB1, 3, 4 and 5) were considered as mono-allelic meaning peptides were most likely bound to HLA class II heterodimers comprising capture beta chains paired with the capture alpha chains. However, the possibility remains that some peptides may have bound to HLA II heterodimers comprising knock-in the beta chains paired with a distinct endogenously expressed alpha chains.

[0672] Conversely, for HLA-DP and HLA-DQ loci, the alpha chains exhibit important allelic variants such that the presence of both knock-in and endogenous alpha chain alleles creates the potential for multiple heterodimers. For example, knock-in alpha and beta chains coding for distinct HLA-DP and HLA-DQ heterodimers can each pair with endogenously expressed alpha and beta chains making up to four unique heterodimers for each HLA-DP and HLA-DQ MAPTAC™ construct. Therefore, binding specificities among the purified MAPTAC™ peptide population are not mono-allelic. To mitigate this endogenous pairing problem, a construct that lacked the alpha chain was used that (sans-alpha knock-ins) enabled us to identify the population of peptides that likely bind to HLA heterodimers comprising endogenously alpha chains and MAPTAC™ beta chains. These peptides were computationally subtracted from the corresponding alpha+beta chain MAPTAC™ experiments to approximate a population of peptides specific to the mono-allelic MAPTAC™ alpha+beta combination.

[0673] Each peptide was assigned to one or more protein-coding transcripts within the UCSC hg19 gene annotation. Since many peptide identifications overlap others and thus constitute mostly redundant information, the peptides were grouped into "nested sets", each meant to correspond to ~1 unique binding event, as shown in FIG. 11C. For instance, the peptides GKAPILIATDVASRGLDV (SEQ ID NO: 16), GKAPILIATDVASRGLD (SEQ ID NO: 17), and KAPILIATDVASRGLDV (SEQ ID NO: 18) all contain the conserved sequence KAPILIATDVASRGLD (SEQ ID NO: 19), and probably all bind MHC in the same register. In order to nest peptides of a given data set, a graph was built in which each node corresponded to a unique peptide, and an edge was created between any pair of peptides sharing at least one 9mer and mappable to at least one common transcript. The clusters command in the R package *igraph* was used to identify clusters of connected nodes, and each cluster was defined as a nested set. This procedure guarantees that any two peptides that meet the edge criteria (≥1 common 9mer and ≥1 common transcript) are placed within the same nested set.

Analysis of previously published MS data

DRB1	DRB3/4/5	Haplotype Freq. (%)
DRB1*03:01	DRB3*01:01	23.2
DRB1*04:01	DRB4*01:01	9
DRB1*04:02	DRB4*01:01	2.2
DRB1*04:04	DRB4*01:01	3.2
DRB1*07:01	DRB4*01:01	13.2
DRB1*09:01	DRB4*01:01	2.8
DRB1*11:01	DRB3*02:02	7.2
DRB1*11:04	DRB3*02:02	2.8
DRB1*13:01	DRB3*02:02	2.6
DRB1*13:02	DRB3*03:01	2.6
DRB1*14:01	DRB3*02:02	2.2
DRB1*15:01	DRB5*01:01	28.6
Note that DRB4*01:01 has been shown to be identical to DRB4*01:03 (ebi.ac.uk/cgi-bin/ipd/imgt/hla/get_allele.cgi?DRB4*0101)		

[0674] For each DRB1/3/4/5 allele present in each (imputed) genotype, twenty peptides from our mono-allelic MAPTAC™ data were spiked in. These augmented datasets were then submitted to GibbsCluster-v2.0.

Characterizing observed cleavage sites of MHC II peptides

[0675] Disclosed herein is a large dataset of naturally processed and presented peptides MHC II peptides by merging peptide identifications across several studies that used immunopurification to profile human tissues (**Table 2**). Since many peptides share the same N-terminus (e.g. GKAPILIATDVASRGLDV (SEQ ID NO: 16) and GKAPILIATDVASRGLD (SEQ ID NO: 17)) or the same C-terminus (e.g. GKAPILIATDVASRGLD (SEQ ID NO: 17) and KAPILIATDVASRGLD (SEQ ID NO: 19)), two sets of non-redundant cut sites were curated, one for N-termini and one for C-termini. Then, an equivalent number of unique non-observed N-terminal and C-terminal cut sites were sampled at random from the set of genes that had produced at least one MHC II peptide. These four data sets were referred to as N-terminal hits, C-terminal hits, N-terminal decoys, and C-terminal decoys. In addition, a naming system was used to refer to positions upstream of peptides, within peptides, and downstream of peptides which is shown in **FIG. 41**.

[0676] The frequency of each amino acid was determined for positions U10 through N3 for N-terminal hits, and these frequencies were compared to those observed for N-terminal decoys. To determine whether hits and decoys showed a significant difference in the rate of a given amino acid at a given position, a 2x2 table (e.g. count of hits for which U1 is lysine, count of decoys for which U1 is lysine; count of hits for which U1 is not lysine, and count of decoys for which U1 is not lysine) was created and scored by a Chi-square test. An analogous approach was use for analyzing amino acid frequencies in positions C3 through D10 of C-terminal hits and decoys.

[0677] A second analysis considered statistical linkages between residues immediately preceding and following cleavage events. First, the count of U1:N1 pairs (A:A, A:C, A:D, …, Y:V, Y:W, Y:Y) was compared for N-terminal hits vs. N-terminal decoys, and significance of enrichment/depletion for each pair was determined by a Chi-square test of a 2x2 contingency table (e.g. count of hits with P:K, count of decoys with

peptides GKAPILIATDVASRGLDV (SEQ ID NO: 16), GKAPILIATDVASRGLD (SEQ ID NO: 17), and KAPILIATDVASRGLDV (SEQ ID NO: 18) counted as a single observation).

[0697] Many samples from the ovarian study had corresponding RNA-Seq data, but some did not. In those cases, expression was estimated averaging across the samples with available RNA-Seq data. For the colorectal and melanoma studies, there was no corresponding RNA-Seq for any samples, so averages were calculated across surrogate samples using data from GTEx and TCGA. In all cases, raw fastqs were obtained and aligned and quantified them according to the same protocols as described above for the ovarian study's RNA-Seq.

[0698] Two matrices were created representing expected and observed counts, referred to as **E** and **O**, respectively, wherein rows correspond to genes and columns correspond to samples. The matrix **E** was first populated by multiplying each gene's length by its expression in each sample; then the columns of **E** were rescaled to make the column sums of **E** match the column sums of **O**. Finally, analysis was made at the gene level by comparing the row sums of **E** to the row sums of **O**. Genes were highlighted according to their presence and concentration in human plasma.

Analysis of genes related to autophagy

[0699] Two autophagy-related gene sets were defined. The first set comprised proteins experimentally identified as physical interaction partners of known autophagy-related genes. For each canonical autophagy-related gene (genenames.org/cgi-bin/genefamilies/set/1022) used as bait in an IP-MS experiment deposited in the Autophagy Interaction Network data base (accessed from besra.hms.harvard.edu/ipmsmsdbs/cgi-bin/downloads.cgi), the top 100 protein identifications according to the "WD" confidence score (besra.hms.harvard.edu/ipmsmsdbs/cgi-bin/tutorial.cgi) were identified. Pooling across 22 experiments, a set of 1004 unique genes were obtained confidently associated with at least one canonical autophagy-related gene. (FIG. 18C)

[0700] A second set of autophagy-related genes were identified using a study that measured pan-proteome protein abundance in baby mouse kidney epithelial (iBMK) cells pre- and post- ATG5 knockout using SILAC (sciencedirect.com/science/article/pii/S1097276514006121). Genes with a t-statistic >5 were classified as being stabilized by ATG5 knockout (pre-starvation conditions; variable "Intercept_t" in supplemental data file mmc2.xls). To map each mouse Uniprot ID to an hg19 UCSC ID, the human UCSC protein sequence was determined with which the mouse Uniprot sequence shared the most 9mers. (FIG. 18C)

[0701] Based on **FIG. 18B**, the ratio R of observed to expected peptides per gene was calculated, adding a pseudocount of one to both the numerator and the denominator, e.g. R=(O+1)/(E+1). Log(R) was taken to represent the relative enrichment (Log(R)>0) or depletion (Log(R)<0) of each gene felt that there was insufficient information to quantify relative enrichment/depletion for these genes. Among the genes with valid Log(R) calculations, Log(R) distributions were plotted for those in the IP-MS dataset, those in the SILAC dataset, and those in neither dataset (**FIG. 18C**).

Analysis of source gene localization, related to FIG. 18D

properties of the donor's cells. The 1:499 ratio is not far from what would be used if down-sampling was not implemented.

[0715] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

EXAMPLE 11: High-throughput identification and validation of HLA class II allele binding epitopes

[0716] In this example, a representative reliable, high-throughput method using time resolved fluorescence energy transfer (TR-FRET) for identification and validation of novel MHC-II allele-binding peptides is described. The assay has several parts, (1) transfecting cells with a vector construct suitable for expressing and secreting MHC-II α and β chains having a fluorescence tag for the FRET assay, (2) purifying the secreted MHC-II construct protein products, (3) performing a peptide exchange assay (**FIG. 22A)**. **FIG. 22B** and **FIG. 23** further exemplify the design and the procedure. The assay as described herein promotes fast and efficient detection and validation protocol, as it may not require stable cell lines, and encompasses simple isolation strategies. In addition, the tetramer or multimer can be used to detect antigen-specific CD4 cells, for example, after *neonmhc2* predicted epitopes are administered in vivo, and the immune response generated thereafter is used to verify CD4+ T cell response.

CLIP-TR-FRET Assay for identifying high affinity MHC Class-II binding peptides

[0717] Presented herein are exemplary vectors for expression of HLA class II α and β chains driven by a CMV promoter in a single construct, the protein product of which yields a properly folded α and β chain pairs. In a properly folded α and β chain form, the α1 subunit and the β1 subunit are in dimer form, the α1 subunit and the β1 subunits forming the open accepting end, capable of accepting a peptide, resembling physiological configurations. For the purpose of this assay, these vector expressed HLA protein products with the properly folded α and β chain form are called HLA monomers. The expression construct comprises a linker, one or more peptide cleavage sites, secretion signal, dimerization factors, for example c-Fos and Jun, linked with a biotinylation motif (BAP) and a 10X-His-Tag (SEQ ID NO: 20). A placeholder peptide is used to stabilize the monomers and help in secretion. A placeholder peptide can be a CMV peptide. A placeholder peptide can be a CLIP peptide. A

placeholder peptide can be a peptide identified via MS based ligandome for the alleles. A placeholder peptide can be bound covalently to the HLA peptides at the open α1-β1 peptide accepting end.

[0718] An exemplary construct used herein encodes a CLIP placeholder peptide with a thrombin cleavage moiety placed between the CLIP and the β chain, as shown in **FIG. 23** (upper panel). Upon transfection and culture of the transfected cells, such that they reach optimal growth, the cell culture supernatant (medium) comprising the secreted proteins (monomers) were collected and passed through nickel (Ni^{2+}) columns for purification (**FIG. 24**). Expression levels and purification was examined by Coomassie staining. The 28-mer CLIP peptide remains associated with the β chain, which is cleaved by treatment with thrombin (FIG. 24) and thereafter may be dislodged by competing with test (*e.g.*, candidate) HLA-Class-II binding peptide. A test peptide that successfully dislodges the CLIP peptide is accountably a cognate peptide for binding to the MHC-II heterodimer of the construct, based on ability to displace the CLIP peptide as measured by its IC50. A *de novo* test peptide could be used for a competitive displacement reaction as described above can then be identified by mass spectrometry (MS).

[0719] A large collection of HLA-DR heterodimer constructs were made with CLIP placeholder peptides which were successfully secreted and peptide exchange assays were performed.

[0720] It was observed that the peptide placeholder CLIP, derived from CD74, has significant effect on the secretion of HLA class II monomers. The edited canonical CLIP peptide having the CD74 sequence PVSKMRMATPLLMQA (SEQ ID NO: 1) (designated as CLIP0 in FIG. 25A-25C) was generally used as placeholder sequence. However, it was seen that some HLA-DR peptides, for example, DRB1*12:01 and DRB1*13:02 had low yield with the canonical peptide (**Table 6**). It was observed that certain HLA DRB allelic dimers have a binding sequence longer sequence covering the whole or parts of the amino acids in the sequence: LPKPPKPVSKMRMATPLLMQALPM (SEQ ID NO: 21) (CLIP1) (**FIG. 25A**). Indeed, using CLIP1 sequence instead of CLIP0 sequence in case of DRB1*12:01 and DRB1*13:02 improved the secretion yield of the HLA dimers (**FIG. 25B-25C**).

Table 6

DRB1*	**Peptide placeholder**	**SEQ ID NO:**	**Yield (mg/L)**
01:01	PVSKMRMATPLLMQA (**CLIP**)	1	20
	LPLKMLNIPSINVH (**CMV**)	22	100
	PKYVKQNTLKLAT (**HA**)	23	85
12:01	PVSKMRMATPLLMQA (**CLIP**)	1	<3
13:02	PVSKMRMATPLLMQA (**CLIP**)	1	<3

De novo Screen of Peptides by Successful Peptide Exchange Assay using STII-TR-FRET

[0721] Peptides can be screened *de novo* using the assay involving expressing HLA-monomer proteins described above in cell lines, such as Expi293 cells, collected and purified from the supernatant, and subjected to peptide exchange assay. HLA class II binding peptides predicted by the prediction algorithms were tested using peptide exchange assay. Peptides exchange assay can be performed using a method involving fluorescence polarization. For example, any fluorophore can be used to label either the placeholder peptide, or to label the

P-191	5167	-----	----	---	
P-192	>40,000				No displacement up to 40µM
HA	23	23	23	0	
ASP51	>40,000	>40,000	N/A	N/A	No displacement up to 40µM.

Peptide exchange validation using Differential Scanning Fluorometry (DSF)

[0723] In this method a high throughput assay for screening peptides that can bind to a particular HLA allele and also, the intensity of the peptide binding to the HLA dimer is determined (**FIG. 26F**). In this assay a fluorescent probe is used, which binds to the hydrophobic residues of a protein and therefore can bind to the MHC alleles, only when the alleles dissociate from each other by application of heat. When an MHC class II dimer binds a cognate peptide, the dimers are held together in its dimeric form. When heat is applied to an MHC dimer-peptide in bound form, the weak binding peptides dissociate faster from the MHC class II protein dimer, allowing the fluorophore to bind to the dissociated MHC alpha and beta chains and producing high fluorescence. Fluorescence is recorded as a function of temperature. Representative melting curves are shown in **FIG. 26F**. Melting curves can be compared to determine the strong binders (fluorescence detected at higher temperature) from weak binders (fluorescence detected at lower temperature).

[0724] *Use of soluble HLA-DM (HLA-sDM) as a catalyst for MHC class II peptide exchange:* HLA-DM is a natural chaperone and peptide exchange catalyst for HLA-DR, -DP, and -DQ molecules. It is an integral membrane protein and occurs as a heterodimer of alpha and beta polypeptide chains (DMA and DMB). Peptide exchange as described in this section is performed using a soluble form of HLA-DM (e.g., HLA-sDM protein) as chaperon for the HLA-DR, -DP, and -DQ exchanges. HLA-sDM protein is produced via a transient transfection in Expi-CHO cells as shown in **FIG. 26G**. Briefly, a recombinant HLA-sDM construct is designed, as shown graphically in FIG. 26G upper half. The recombinant HLA-sDM construct comprises a CMV constitutive promoter, upstream of a leader sequence and operably linked with the promoter. The leader sequence helps in the secretion of the product (secretion signal). At the 3'-end of the leader sequence a coding sequence for HLA-DM beta chain ectodomain (and lacks a transmembrane domain) is introduced. A sequence encoding a biotinylation motif (BAP) is ligated 3'of the beta chain-encoding sequence. A sequence encoding the HLA-DM alpha chain ectodomain (and lacks a transmembrane domain) is placed with a secretion sequence (leader) at its 5' end, separated from the BAP sequence by an intervening ribosomal skipping sequence. The HLA-DM alpha chain sequence is ligated at the 3' end with a 10X HIS-tag (SEQ ID NO: 20). Once formed the heterodimeric HLA-sDM is secreted outside the cell. When this construct is expressed in Expi-CHO cells the HLA-sDM protein is secreted into the medium culture medium.

[0725] Expi-CHO cells were transfected with a plasmid vector expressing the HLA-sDM construct, and cultured over a period of about 14 days. The protein was secreted into the culture medium over the period of culture. The HLA-sDM protein was purified from the culture in a process very similar to purifying MHC-II proteins. MHC-II peptide exchange can be performed efficiently with acid and HLA-sDM, or without acid, and with octyl glucoside. Size exclusion chromatography was performed to assess peptide exchange, results were

of detergent (1% Octyl glucoside vs 1.6% NP40 was evaluated in membrane permeabilization for harvesting full length MHCII alleles. In some examples, effect of temperature, or the probe used, or the purification methods or the target format were individually evaluated (**FIG. 28B**).

[0733] Effect of purification method using either conformation specific antibody L243 or His-tag purification were evaluated. The results are shown in **FIG. 28D**. Each data point is depicted in the table on the left and is represented as a single dot in the graph on the right. The dots in the graph align roughly along a 45-degree angle to either axes, and with a r value of 0.9621, which indicate that the IC50 values from both the purification methods are in agreement with each other. It also shows that the rank order of peptide potencies does not change between the purification methods.

[0734] Effect of the choice of the HLA class II proteins in soluble form (sDR1) versus the full-length form (fDR1) was evaluated and **FIG. 28E** shows that choice of target format does not affect the peptide potency. Shown on the left are average IC50 values from experiments using sDR1 form or fDR1. These data are plotted to obtain the graph on the right hand side. The data points, each depicted by a single dot, align roughly along a 45 degree angle to either axes, and with a r value of 0.9365, which indicate that the IC50 values from both the forms used correlate well with each other. It also shows that the rank order of peptide potencies does not change between the purification methods.

[0735] **FIG. 28F** shows a graphical view of an exemplary evaluation method of *neonmhc2* and NetMHCIIpan predicted peptides in binding assay and identification of discordant peptides. Fluorescence polarization assay was used to evaluate the *Neonmhc2* and NetMHCIIpan predicted peptides in actual peptide binding assays. For the assay, 60 nM of thrombin digested soluble HLA-DRB1*15:01 was incubated with an FITC-tagged super binder probe peptide (PVVHFFK(FITC)NIVTPRTPPY (SEQ ID NO: 24)) (10 nM per assay) and the assay peptide for 5 hours at 37°C in an assay buffer (pH 5.2). Fluorescence polarization was examined from which, % probe displacement was calculated. As shown in **FIG. 28G**, inhibition of the super binder fluorescent peptide was proportional to the concentration of the predicted peptide indicating good specificity of the assay. Striking differences are seen between the performances of the *Neonmhc2* and the NetMHCIIpan predicted peptides. With *Neonmhc2* predicted peptides more peptides were positively bound, and with higher degree of inhibition; whereas the NetMHCIIpan predicted peptides were overall poor performers in comparison to the *Neonmhc2* peptides. **FIG. 28H** summarizes the evaluation of *Neonmhc2* predicted peptides in binding assay. As depicted by the pie charts, of the double negative peptides (peptides that were not predicted by any of the NetMHCIIpan or *Neonmhc2*) only 5% turned out to be binders and 95% non- binders. Of the NetMHCII predicted peptides, 40% were binders by fluorescence polarization detection of probe displacement assay, and of the *Neonmhc2* predicted peptides to be binders, 100% were found to be true binders by the probe displacement assay.

[0736] FITC-labelled probes were prepared by reviewing previously published peptides shown by Sette et al., to bind specific alleles. These peptide sequences were then analyzed using predicted class II binding core to identify the minimal 9-mer core of the peptide and the anchor residues. This information was then considered when selecting a residue position for lysine substitution and FITC labelling. For example, in the table below (**Table 10**) the sequences as described in Sette et al. (Sidney J, Southwood S, Moore C, et al. Measurement of

MHC/peptide interactions by gel filtration or monoclonal antibody capture. *Curr Protoc Immunol*. 2013;Chapter 18:Unit–18.3. doi:10.1002/0471142735.im1803s100) (hereinafter "Sette's Sequences")) are listed. The predicted class II binding core for each peptide were underlined in the context of a specific allele. The bold font denotes anchor positions that were identified as a result of epitope improvement. In some cases, the same peptide sequence can be used for different alleles.

Table 10

Shorthand ID	**Allele**	**Sette's Sequences**	**SEQ ID NO:**	**Probe Sequences Selected**	**SEQ ID NO:**
SB-DR7/11	DRB1*07:01	YATF**F**IK**A**N**S**KF**I**GITE	25	YATFFI***K***ANSKFIGITE	25
SB-DR7/11	DRB1*11:01	YATFF**I**KAN**S**K**K**FIGITE	25	YATFFI***K***ANSKFIGITE	25
SB-DR9	DRB1*09:01	TLSVT**F**IG**A**APLILSY	26	TLSVTFIGAAP***K***ILSY	29
SB-DR4/15	DRB1*15:01	PV**V**HF**F**KN**I**V**T**PRTPPY	27	PVVHFF***K***NIVTPRTPPY	27
SB-DR4/15	DRB1*04:01	PVVHF**F**KN**I**V**T**PR**T**PPY	27	PVVHFF***K***NIVTPRTPPY	27
SB-DR3	DRB1*03:01	YAR**I**RR**D**G**C**LL**R**LVD	28	YARI***K***RDGCLLRLVD	30

[0737] Based on positioning as described above, an internal lysine for FITC conjugation was chosen by focusing on positions within the binding core (underlined) that are not in blue –highlighted in red were these positions as appropriate positions for FITC conjugation. For sequences that did not have an internal lysine for FITC conjugation, a manual approach was undertaken where a comparison to an allele's binding motif to the peptide sequence was performed, and a position for internal lysine substitution was selected for the DRB1*09:01 and DRB1*03:01 peptides (see above table). More specifically, a leucine residue for DRB1*09:01, and an arginine residue for DRB1*03:01 were substituted with lysines to allow for FITC conjugation. This substitution strategy was based on the MAPTAC-derived motifs, where manual identification of positions with no strong amino acid preference (also in the middle of the *neonmhc2* predicted 9-mer core) because the conjugated fluorophore may be more likely to emit polarized light when bound (i.e., more restricted motion of the fluorophore).

EXAMPLE 12: HLA class II binding and processing rules for identifying therapeutically targetable cancer antigens

[0738] Increasing evidence indicates CD4+ T cells can recognize cancer-specific antigens and control tumor growth. However, it remains difficult to predict the antigens that will be presented by human leukocyte antigen class II molecules (HLA class II) - hindering efforts to optimally target them therapeutically. Obstacles include inaccurate peptide-binding prediction and unsolved complexities of the HLA class II pathway. In this Example, an improved technology for discovering HLA class II binding motifs is described. Further, described herein is a comprehensive analysis of tumor-ligandomes conducted to learn processing rules relevant in the tumor microenvironment (TME).

[0739] 40 HLA class II alleles were profiled and it was shown that binding motifs are highly sensitive to HLA-DM, a peptide loading chaperone. The intratumoral HLA class II presentation was revealed to be dominated by

Spike-in peptides for deconvolution analysis

[0790] An exemplary list of 20 example peptides used per allele in the spike-in analysis. Peptides were selected by requiring a minimum SPI of 70, length between 12 and 20 amino acids, and by not allowing a 9mer overlap with any binders observed for other MAPTAC™ -profiled DR alleles. Additionally, no two spike-in peptides for a given allele share a 9mer. See for example, at least FIGs 35, 36A-36C, among others.

Collated TGEM data set for selected alleles Supplemental Experimental Procedures

[0791] HLA class II tetramer results for DRB1*01:01, DRB1*03:01, DRB1*04:01, DRB1*07:01, DRB1*11:01, and DRB1*15:01 for diverse pathogen and allergen peptides and their corresponding NetMHCIIpan and *neonmhc2* predictions. Data were curated from papers published by Kwok and colleagues. See FIG. 38A-38C, among others.

Supplemental Methods

HLA class II allele frequencies and affinity data statistics, related to FIG. 12A and 12E

[0792] Allele frequencies were obtained from resource, bioinformatics.bethematchclinical.org/hla-resources/haplotype-frequencies/high-resolution-hla-alleles-and-haplotypes-in-the-us-population. The mhc_ligand_full.csv dataset was downloaded from IEDB data (iedb.org/database_export_v3.php) on September 21, 2018. Valid affinity measurements were required to have a "Method/Technique" equal to "cellular MHC/competitive/fluorescence", "cellular MHC/competitive/radioactivity", "cellular MHC/direct/fluorescence", "purified MHC/competitive/fluorescence", "purified MHC/competitive/radioactivity", or "purified MHC/direct/fluorescence" and an "Assay Group" equal to "dissociation constant KD", "dissociation constant KD (~EC50)", "dissociation constant KD (~IC50)", "half maximal effective concentration (EC50)", or "half maximal inhibitory concentration (IC50)". A measurement was attributed to the Søren Buus group (University of Copenhagen, Denmark) if the string "Buus" appeared in the "Authors" field. Otherwise, if the authors field included the strings "Sette" or "Sidney", a measurement was attributed to the Alessandro Sette group (La Jolla Institute for Immunology, U.S.A). All other measurements were labeled as "Other". For the purposes of enumerating strong binders, only peptides with a measured affinity stronger than 100 nM were counted

MAPTAC™ protocol overview, related to FIG. 2: DNA construct design

[0793] The gene sequences for HLA class I and HLA class II alleles were identified by the IPD-IMGT/HLA webpage (ebi.ac.uk/ipd/imgt/hla) and used to design recombinant expression constructs. For HLA class I, the α-chain was fused with a C-terminal GSGGSGGSAGG linker (SEQ ID NO: 10), followed by the biotin-acceptor-peptide (BAP) tag sequence GLNDIFEAQKIEWHE (SEQ ID NO: 11), a stop codon, and a variable DNA barcode, and cloned into the pSF Lenti vector (Oxford Genetics, Oxford, UK) via the NcoI and XbaI restriction sites. The HLA class II constructs (DR, DP and DQ) were similarly cloned into pSF Lenti via the NcoI and XbaI restriction sites and consisted of the β-chain sequence fused on the C-terminus to the linker-BAP sequence from the HLA class I construct (SGGSGGSAGGGLNDIFEAQKIEWHE (SEQ ID NO: 12)), followed by another short GSG linker an a F2A ribosomal skipping sequence (VKQTLNFDLLKLAGDVESNPGP (SEQ ID NO: 13)), the sequence of the α-chain, an HA tag (GSYPYDVPDYA (SEQ ID NO: 14)), a stop codon, and a variable DNA barcode. For all DR alleles the beta-chain was paired with DRA*01:01. The

25 passages. K562 cells and KG-1 cells (ATCC, Manassas, VA) were grown in IMDM (Thermo Fisher Scientific) media plus 10% heat inactivated FBS, 1% penicillin/streptomycin, 1% sodium pyruvate, and 1% MEM-NEAA. Cells were cultured twice weekly and discarded after 25 passages.

[0796] Lentivirus for transduction of B721.221, KG-1, and K562 cells were produced in HEK293T cells grown to 80% confluency. Six micrograms of the genome vector psFLenti encoding HLA class I or HLA class II (described in previous sections) was mixed with 5.3ug of the lentivirus packaging vector psPAX2 and 1.81ug of the envelope vector pMD.2. DNA was mixed with Opti-MEM (Thermo Fisher Scientific) and the transfection reagent, Fugene HD (Promega, Madison, WI), and the mixture was incubated at room temperature for 15 minutes. The mixture was then added dropwise onto the dish of HEK293T cells and incubated for 72 hours. Supernatant was then harvested, and lentiviral titers were tested using Lenti-X GoStix (Takara Bio Inc., Japan). For transduction, cells were seeded in 12-well flat bottom plates (Corning Inc., Corning, NY) and mixed with lentiviral supernatant with 6ug/ml polybrene (Sigma-Aldrich). Cells mixed with lentivirus were spun at 32°C at 800xg for 90 minutes. Cells were resuspended in warm media and incubated in a 37°C incubator at 5% CO_2 for 72 hours. Cells were then selected using 1ug/ml puromycin for 2 weeks. After selection, at least 50 million cells were harvested, centrifuged, washed once with 1 x Gibco DPBS (Thermo Scientific), and flash frozen in liquid nitrogen for mass spectrometric analysis.

BirA protein expression and purification

[0797] The pET19 vector encoding *E. coli* BirA fused to a C-terminal hexa-histidine tag (SEQ ID NO: 15) was used. Chemical competent *E. coli* BL21 (DE3) cells (New England Biolabs) were transformed with a BirA expression plasmid (pET19 vector encoding *E. coli* BirA fused to a C-terminal hexa-histidine (SEQ ID NO: 15)), grown at 37°C in LB broth plus 100 μg/ml ampicillin to an OD_{600} of 0.6–0.8 and cooled to 30°C before expression was induced by adding 0.4 mM isopropyl-β-D-thiogalactopyranoside. *E. coli* cell growth continued at 30°C for 4 h. *E. coli* cells were harvested by centrifugation at 8000 x g for 30 minutes at 4 °C and stored at –80°C until use. Frozen cell pellets expressing recombinant BirA were resuspended in IMAC buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl) with 5 mM Imidazole, incubated with 1 mg/ml lysozyme for 20 minutes on ice and the lysed by sonication. Cellular debris and insoluble materials were removed by centrifugation at 16,000 x g for 30 minutes at 4°C. The cleared supernatant was subsequently loaded on a HisTrap HP 5 mL column using the ÄKTA pure chromatography system (GE Healthcare), washed with IMAC buffer plus 25 mM and 50 mM imidazole before elution with 500 mM imidazole. Fractions containing BirA were pooled and dialyzed against 20 mM Tris-HCl pH 8.0 with 25 mM NaCl and were loaded on a HiTrap Q HP 5mL column (GE Healthcare, Chicago, IL) and eluted by applying a linear gradient from 25 to 600 mM NaCl. Fractions containing highly pure BirA were pooled, buffer exchanged in storage buffer (20 mM Tris-HCl pH 8.0 100 mM NaCl, 5% glycerol) and concentrated to around 5-10 mg/mL, aliquoted, and flash frozen in liquid nitrogen for storage at –80°C. BirA protein concentration was determined by UV spectroscopy at $OD_{280\ nm}$ using a calculated extinction coefficient of $\varepsilon = 47{,}440\ M^{-1}\ cm^{-1}$.

Western blotting protocol

Mapping peptides to genes and "nested sets", related to FIGs. 30A- 30C, **FIGs. 31A-31D**, and **FIGs 32A-32E**

[0809] Each peptide was assigned to one or more protein-coding transcripts within the UCSC hg19 gene annotation (genome.ucsc.edu/cgi-bin/hgTables). Since many peptide identifications overlap others and thus constitute mostly redundant information, peptides were grouped into "nested sets", each meant to correspond to ~1 unique binding event. For instance, the peptides GKAPILIATDVASRGLDV (SEQ ID NO: 16), GKAPILIATDVASRGLD (SEQ ID NO: 17), and KAPILIATDVASRGLDV (SEQ ID NO: 18) all contain the conserved sequence KAPILIATDVASRGLD (SEQ ID NO: 19), and probably all bind MHC in the same register. In order to nest peptides of a given data set, a graph was built in which each node corresponded to a unique peptide, and an edge was created between any pair of peptides sharing at least one 9mer and mappable to at least one common transcript. The clusters command in the R package *igraph* (Team, 2014) (cran.r-project.org/web/packages/igraph/citation.html) was used to identify clusters of connected nodes, and each cluster was defined as a nested set. This procedure guarantees that any two peptides that meet the edge criteria (≥1 common 9mer and ≥1 common transcript) are placed within the same nested set. The nests were used for sequence logo generation (logos were generated using the shortest peptide in each nested set; **FIG. 30A-30C**, machine learning (importance weights across peptides in a nested set sum to one; **FIG. 31A-31D**), and the gene bias analysis (each nested set was counted as one observation rather than each individual peptide; **FIG. 32A-32E**).

Analysis of amino acid frequencies, related to FIG. 12F

[0810] Amino acid frequencies in the human proteome were calculated based on sequences for all protein-coding genes in the UCSC hg19 annotation (selecting one transcript at random for genes represented by multiple transcript isoforms). IEDB frequencies were determined by identifying the unique set of peptides with at least one affinity observation ≤100 nM (excluding peptides with hexavalent polyhistidine at their C-terminus). MAPTAC™ frequencies were first considered in the context of the standard forward-phase protocol across five DRB1 alleles (DRB1*01:01, DRB1*03:01, DRB1*09:01, and DRB1*11:01), using only one peptide (the longest) per nested set. In addition, MAPTAC™ frequencies were separately calculated for the subset of samples processed by the reduction and alkylation protocol. MS data from external datasets were analyzed without respect to potential allele of origin and likewise using the longest peptide per nested set.

Building HLA class I sequence logos, related to FIG. 37B

[0811] For each HLA class I allele, a length-9 sequence logo was created by profiling amino acid frequencies in the first five positions (mapping to logo positions 1-5) and last four positions (mapping to logo positions 6-9) of corresponding peptides. In this manner, peptides contributed to the sequence logo regardless of their length. As in the HLA class II logos, letter heights are proportional to the frequency of each amino acid in each position, and color coding is used for amino acids with frequency ≥10%.

Assessing the performance of HLA class II peptide deconvolution, related to FIG. 30B

[0812] To assess the ability the GibbsCluster (v2.0) tool to cluster multi-allelic HLA class II peptide data by allele of origin, its performance on eight samples were analyzed, including 4 PBMC samples, 1 melanoma cell line (A375), and 3 previously published lymphoblastoid cell lines. For each DRB1/3/4/5 allele present in each

attributable to the tumor (**FIG. 19B**) was estimated as min(1,2*f*), where *f* is the fraction of RNA-Seq reads in CIITA, CD74, or CTSS exhibiting a mutation.

Identification of over- and under-represented genes, related to FIG. 32A and FIG. 39B

[0843] Samples were analyzed from previously published MS experiments that profiled the MHC-II ligandomes of ovarian cancer, colorectal cancer, and melanoma. Many samples from the ovarian cancer dataset had available RNA-Seq; data for these samples was downloaded from SRA (NCBI BioProject PRJNA398141) and aligned to the UCSC hg19 transcriptome using STAR aligner. For ovarian samples that did not have available RNA-Seq, expression was estimated averaging across the samples with available RNA-Seq. For the colorectal and melanoma studies, there was no corresponding RNA-Seq for any samples, so averages were calculated across surrogate samples using data from TCGA (The Cancer Genome Atlas Network). Transcript level gene quantification was performed using transcripts per million (TPM) as calculated by RSEM version-1.2.31. The expression estimates were further processed by summing to the gene level, dropping non-coding genes, and renormalizing such that the total TPM summed to 1000000 (renormalizing across protein-coding genes accounts for library-to-library variation in ncRNA abundance).

[0844] To identify genes over- and under-represented in HLA class II ligandomes, it was analyzed the same three datasets used in the expression analysis. For each gene, our baseline assumption was that it should yield peptides in proportion to its length multiplied by its expression level. To determine the length of each gene, the unique 9mers across all transcript isoforms were enumerated. Gene-level expression was obtained by summing across transcript isoforms. The observed number of peptides mapping to each gene was determined at the nested set level (e.g. peptides GKAPILIATDVASRGLDV (SEQ ID NO: 16), GKAPILIATDVASRGLD (SEQ ID NO: 17), and KAPILIATDVASRGLDV (SEQ ID NO: 18) counted as a single observation).

[0845] Two matrices were created representing expected and observed counts, referred to as **E** and **O**, respectively, wherein rows correspond to genes and columns correspond to samples. The values in **O** were determined by counting peptides per sample at the nested set level. The matrix **E** was first populated by multiplying each gene's length by its expression in each sample; then the columns of **E** were rescaled to make the column sums of **E** match the column sums of **O**. Finally, analysis was made at the gene level by comparing the row sums of **E** to the row sums of **O** (**FIG. 32A**). Genes were highlighted according to their presence and concentration in human plasma. An identical approach was used for identifying over- and under-represented gene in HLA class I data, using melanoma, colorectal cancer, and ovarian cancer data from the same set of studies. For the HLA class I analysis, no nesting was applied, but only unique peptides were counted.

Assessment of binding scores in over-represented genes, related to FIG. 39A

[0846] It was observed that many of the over-represented genes were plasma genes. A comprehensive list of serum genes was obtained and the *neonmhc2* binding scores were compared for HLA DR-bound peptides derived from plasma genes with HLA-DR-bound peptides derived from non-serum genes, as well as with length-matched, non-binding (e.g. not observed in MS) peptides sampled from genes that were represented in the immunopeptidome. For genotyped, multi-allelic datasets that had HLA class II peptides profiled with a pan-DR antibody (the same samples analyzed in **FIG. 30B**), peptides with *neonmhc2* for each DR allele that the

sample expressed were scored. The best score output by *neonmhc2* over all expressed alleles was taken as the representative score for each peptide. The data was pooled across all usable datasets, and the distributions of the scores for each category of peptide were visualized with a boxplot.

Analysis of genes related to protein turnover, related to FIG. 32C

[0847] Two gene sets were identified meant to represent proteins whose turnover is regulated by the proteasome. The first gene set comprised genes with at least one observed ubiquitination site in the cell lines KG1, Jurkat, or MM1S. The second set comprised genes whose levels increased upon application of the proteasome inhibitor Bortezomib (BTZ) of a published paper, applying a p-value filter of 0.01 and selecting the 300 genes with the largest upward fold change.

Comparing explanatory power of bulk tumor vs. antigen presenting cell gene expression, related to FIG. 39C

[0848] Four gene expression profiles were created. The first was meant to represent APCs and estimated by averaging cell type-specific profiles from the above-described single-cell RNA-Seq experiments. The average included "macrophages" (from the head and neck study, the lung study, and the melanoma study), "CLEC9A DCs" (from the lung study), and "monoDCs" (from the lung study). The three other expression profiles correspond to bulk tumor profiles from ovarian cancer, colorectal cancer, and melanoma (**Data FIG. 19A**). The ovarian profile was an average of the samples published by Schuster et al, and the other profiles are derived from the five TCGA samples with the highest tumor cellularity, per tumor type, as previously inferred using the "Absolute" algorithm. For each tumor type, the number of peptides (at the nested-set level) per gene were counted and modeled each gene's peptide count as a function of gene length, APC-specific gene expression, and tumor-specific gene expression using linear regression. The output variable and all input variables were transformed via log(x+1). Using the model's parameter estimates, the contribution of tumor was calculated as $\beta_{tumor}/(\beta_{tumor}+\beta_{APC})$, and the contribution of APCs was calculated as $\beta_{APC}/(\beta_{tumor}+\beta_{APC})$. For each sample, bootstrap re-sampling (M=100) at the gene level was used to calculate confidence intervals for the explanatory proportions.

Characterizing observed cleavage sites of HLA class II peptides, related to FIG. 40A

[0849] Naturally processed and presented HLA class II peptides were analyzed from six datasets: PBMC draws, the DC-like MUTZ3 cell line, colorectal cancer tissue, melanoma, ovarian cancer, and the expi293 cell line. Since many peptides share the same N-terminus (e.g. GKAPILIATDVASRGLDV (SEQ ID NO: 16) and GKAPILIATDVASRGLD (SEQ ID NO: 17)) or the same C-terminus (e.g. GKAPILIATDVASRGLD (SEQ ID NO: 17) and KAPILIATDVASRGLD (SEQ ID NO: 19)), two sets of non-redundant cut sites were curated, one for N-termini and one for C-termini. The naming system shown in **FIG. 41** was used to refer to positions upstream of peptides, within peptides, and downstream of peptides. Upstream and downstream frequencies (…U1 and D1…), were compared against proteome amino acid frequencies and scored significant deviations via Chi-square test. Peptide positions (N1…C1), were compared against frequencies as observed in MS peptides.

Benchmarking the performance of various HLA class II cleavage predictors, related to FIG. 40B

Sodium Acetate	1M	100mM
Sodium Chloride	5M	50mM
Octyl glucoside	10%	1%
MiliQ water	-	Up to 100 ml

[A]DMSO or peptides with the following sequences were used: PPIDGYPNHPCFEPE (SEQ ID NO: 31) (M230), PQILPYPAPEEAQEN (SEQ ID NO: 32) (M231), PQLRQWWAQGADPLA (SEQ ID NO: 33) (M247), LLRPGQIVAFDSTAQ (SEQ ID NO: 34) (M248) or ASLRSWPSTWAPWAS (SEQ ID NO: 35) (M371.

[0869] The buffer was then exchanged using a PD minitrap G-25 desalting column. Sypro orange dye (Fisher S6651) was diluted to 1000X in 100% DMSO. 50 µL working stock of Sypro orange dye at 100X was prepared in desalting buffer. 2 µL of 100X sypro orange dye and 18 µL of desalted peptide exchanged sample was transferred to wells of a 384 white PCR microplate and mixed. The plate was then covered with a transparent plate sealer and the plate was subjected to the following program in a Roche lightcycler 480: (1) heat to 25 °C, hold for 10 seconds; (2) increase the temperature to 99 °C, read plate 20 times/1 °C (3) bring the temperature down to 25 °C and hold for 10 seconds. Melting temperatures were then calculated. Exemplary results are shown in **Table 18** below.

Table 18

Sample	DT_m (sample T_m-control T_m)	Average T_m	Slope	Initial	Low	Peak
DR15 + M230	21.4	79.7	0.5	2.8	2.0	4.8
DR15 + M231	15.6	73.9	0.3	2.7	1.9	5.8
DR15 + M247	20.1	78.4	0.4	3.0	2.3	5.0
DR15 + M248	5.6	63.9	0.4	1.8	1.4	5.1
DR15 + M371	19.9	78.2	0.4	2.5	1.9	4.9
DR15 + DMSO (Control)	NA	58.3	0.8	2.1	1.7	9.2

CLAIMS

1. A method comprising:

(a) processing amino acid information of a plurality of candidate peptide sequences using a machine learning HLA peptide presentation prediction model to generate a plurality of presentation predictions, wherein each candidate peptide sequence of the plurality of candidate peptide sequences is encoded by a genome or exome of a subject, or a pathogen or a virus in the subject, wherein the plurality of presentation predictions comprises an HLA presentation prediction for each of the plurality of candidate peptide sequences, wherein each HLA presentation prediction is indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells; and

(b) identifying, based at least on the plurality of presentation predictions, a peptide sequence of the plurality of peptide sequences as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 according to a presentation PPV determination method.

2. A method comprising:

(a) processing amino acid information of a plurality of peptide sequences encoded by a genome or exome of a subject using a machine learning HLA peptide binding prediction model to generate a plurality of binding predictions, wherein the plurality of binding predictions comprises an HLA binding prediction for each of the plurality of candidate peptide sequences, each binding prediction indicative of a likelihood that one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given candidate peptide sequence of the plurality of candidate peptide sequences,

wherein the machine learning HLA peptide binding prediction model is trained using training data comprising sequence information of sequences of peptides identified to bind to an HLA class II protein or an HLA class II protein analog; and

(b) identifying, based at least on the plurality of binding predictions, a peptide sequence of the plurality of peptide sequences that has a probability greater than a threshold binding prediction probability value of binding to at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject;

wherein the machine learning HLA peptide binding prediction model has a positive predictive value (PPV) of at least 0.1 according to a binding PPV determination method.

3. The method of claim 2, wherein the machine learning HLA peptide presentation prediction model is trained using training data comprising sequence information of sequences of training peptides identified by mass spectrometry to be presented by an HLA protein expressed in training cells.
4. The method of any one of the preceding claims, wherein the method comprises ranking, based on the presentation predictions, at least two peptides identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.
5. The method of any one of the preceding claims, wherein the method comprises selecting one or more peptides of the two or more ranked peptides.
6. The method of any one of the preceding claims, wherein the method comprises selecting one or more peptides of the plurality that were identified as being presented by at least one of the one or more proteins encoded by a class II HLA allele of a cell of the subject.
7. The method of any one of the preceding claims, wherein the method comprises selecting one or more peptides of two or more peptides ranked based on the presentation predictions.
8. The method of any one of the preceding claims, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.07 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test presentation predictions, each test presentation prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject can present a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 500 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 499 decoy peptide sequences contained within a protein encoded by a genome of an organism, wherein the organism and the subject are the same species, wherein the plurality of test peptide sequences comprises a ratio of 1:499 of the at least one hit peptide sequence to the at least 499 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to be presented by the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.
9. The method of any one of the preceding claims, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.1 when amino acid information of a plurality of test peptide sequences are processed to generate a plurality of test binding predictions, each test binding prediction indicative of a likelihood that the one or more proteins encoded by a class II HLA allele of a cell of the subject binds to a given test peptide sequence of the plurality of test peptide sequences, wherein the plurality of test peptide sequences comprises at least 20 test peptide sequences comprising (i) at least one hit peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells and (ii) at least 19 decoy peptide sequences contained within a protein comprising at least one peptide sequence identified by mass spectrometry to be presented by an HLA protein expressed in cells, such as a single HLA

protein expressed in cells (e.g., mono-allelic cells), wherein the plurality of test peptide sequences comprises a ratio of 1:19 of the at least one hit peptide sequence to the at least 19 decoy peptide sequences and a top percentage of the plurality of test peptide sequences are predicted to bind to the HLA protein expressed in cells by the machine learning HLA peptide presentation prediction model.

10. The method of any one of the preceding claims, wherein no amino acid sequence overlap exist among the at least one hit peptide sequence and the decoy peptide sequences.
11. The method of any one of claims 1-10, wherein the machine learning HLA peptide presentation prediction model has a positive predictive value (PPV) of at least 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29, 0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4, 0.41, 0.42, 0.43, 0.44, 0.45, 0.46, 0.47, 0.48, 0.49, 0.5, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.7, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.8, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98 or 0.99.
12. The method of any one of claims 1 and 3-11, wherein the at least one hit peptide sequence comprises at least 5, 10, 20, 50 or 100 hit peptide sequences.
13. The method of any one of claims 1 and 3-12, wherein the at least 499 decoy peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 decoy peptide sequences.
14. The method of any one of claims 1 and 3-13, wherein the at least 500 test peptide sequences comprises at least 2500, 5000, 10000, 25000, 50000 or 100000 test peptide sequences.
15. The method of any one of claims 1 and 3-14, wherein the top percentage is a top 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1.00% or 2.00%.
16. The method of any one of claims 2-11, wherein the at least one hit peptide sequence comprises at least 5, 10, 20, 50 or 100 hit peptide sequences.
17. The method of any one of claims 2-11 and 16, wherein the at least 19 decoy peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 decoy peptide sequences.
18. The method of any one of claims 2-11, 16 and 17, wherein the at least 20 test peptide sequences comprises at least wherein the at least 500 test peptide sequences comprises at least 500, 1000, 2000, 5000 or 10000 test peptide sequences test peptide sequences.
19. The method of any one of claims 2-11 and 16-18, wherein the top percentage is a top 5%, 6%, 7%, 8%, 9%, 10%, 15%, or 20%.
20. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 11 for the protein encoded by the corresponding HLA allele of Table 11.
21. The method of any one of the preceding claims, wherein the PPV is more than the PPV of column 2 of Table 16 for the protein encoded by an HLA class II allele.
22. The method of any one of the preceding claims, wherein the subject is a single subject.
23. The method of any one of the preceding claims, wherein the cells are cells expressing a single protein encoded by a class II HLA allele of a cell of the subject.

24. The method of any one of claims 1-23, wherein the cells are monoallelic HLA cells, or cells expressing an HLA allele with an affinity tag.
25. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is associated with a cancer.
26. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.
27. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is overexpressed by a cancer cell of the subject.
28. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.
29. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is a cancer cell-specific peptide.
30. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is expressed by a cancer cell of the subject.
31. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.
32. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not encoded by a non-cancer cell of the subject.
33. The method of any one of the preceding claims, wherein at least one peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.
34. The method of any one of the preceding claims, wherein each peptide sequence of the plurality of peptide sequences is not expressed by a non-cancer cell of the subject.
35. The method of any one of the preceding claims, wherein the method comprises obtaining the plurality of peptide sequences of the subject.
36. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject.
37. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject.
38. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject that encodes the plurality of peptide sequences encoded by a genome or exome of a subject by a computer processor.
39. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by genomic or exome sequencing.
40. The method of any one of the preceding claims, wherein the method comprises obtaining a plurality of polynucleotide sequences of the subject by whole genome sequencing or whole exome sequencing.

41. The method of any one of the preceding claims, wherein the method further comprises formulating a composition comprising one or more of the selected sub-set of peptide sequences.

42. The method of any one of the preceding claims, wherein the HLA class II protein comprises an HLA-DR, HLA-DQ, or an HLA-DP protein.

43. The method of any one of the preceding claims, wherein the HLA class II protein is selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01, HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02, HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01 and HLA-DRB5*01:01; an HLA-DR protein selected from the group consisting of: DPA*01:03/DPB*04:01, DRB1*01:01, DRB1*01:02, DRB1*03:01, DRB1*04:01, DRB1*04:02, DRB1*04:04, DRB1*04:05, DRB1*07:01, DRB1*08:01, DRB1*08:02, DRB1*08:03, DRB1*09:01, DRB1*11:01, DRB1*11:02, DRB1*11:04, DRB1*12:01, DRB1*13:01, DRB1*13:02, DRB1*13:03, DRB1*14:01, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*16:02, DRB3*01:01, DRB3*02:01, DRB3*02:02, DRB3*03:01, DRB4*01:01, DRB4*01:03 and DRB5*01:01; an HLA-DP protein selected from the group consisting of: DPB1*01:01, DPB1*02:01, DPB1*02:02, DPB1*03:01, DPB1*04:01, DPB1*04:02, DPB1*05:01, DPB1*06:01, DPB1*11:01, DPB1*13:01, DPB1*17:01; or an HLA-DQ protein complex selected from the group consisting of: A1*01:01+B1*05:01, A1*01:02+B1*06:02, A1*01:02+B1*06:04, A1*01:03+B1*06:03, A1*02:01+B1*02:02, A1*02:01+B1*03:03, A1*03:01+B1*03:02, A1*03:03+B1*03:01, A1*05:01+B1*02:01 and A1*05:05+B1*03:01.

44. The method of any one of the preceding claims, wherein the peptides presented by the HLA protein have a length of from 15-40 amino acids.

45. The method of any one of the preceding claims, wherein the personalized cancer therapy further comprises an adjuvant.

46. The method of any one of the preceding claims, wherein the personalized cancer therapy further comprises an immune checkpoint inhibitor.

47. A method for identifying HLA class II specific peptides for immunotherapy for a subject, comprising:

(a) obtaining, by a computer processor, a candidate peptide comprising an epitope, and a plurality of peptide sequences, each comprising the epitope;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences to an immune cell, each presentation prediction indicative of a likelihood that one or more proteins encoded by an HLA class II allele can present a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by an HLA protein expressed in cells and identified by mass spectrometry;

(c) selecting a protein from the one or more proteins encoded by the HLA class II allele of a cell of the subject, predicted to bind to the candidate peptide by the machine-learning HLA-peptide presentation prediction model, wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the candidate peptide to an immune cell;

(d) contacting the candidate peptide with the selected protein such that the candidate peptide competes with a placeholder peptide associated with the selected protein; and

(e) identifying the candidate peptide as a peptide for immunotherapy specific for the selected protein based on whether the candidate peptide displaces the placeholder peptide.

48. A method for assaying immunogenicity of a MHC class II binding peptide, comprising:

(a) selecting a protein encoded by an HLA class II allele predicted by a machine-learning HLA-peptide presentation prediction model to bind to the MHC class II binding peptide, wherein the machine-learning HLA-peptide presentation prediction model is configured to generate a presentation prediction for a given peptide sequence, the presentation prediction indicative of a likelihood that one or more proteins encoded by the HLA class II allele can present the given peptide sequence, and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the MHC class II binding peptide;

(b) contacting the peptide with the selected protein such that the peptide competes with a placeholder peptide associated with the selected protein, and displaces the placeholder peptide, thereby forming a complex comprising the HLA class II protein and the MHC class II binding peptide;

(c) contacting the complex with a CD4+ T cell, and

(d) assaying for one or more of activation parameters of the CD4+T cell, selected from the group consisting of: induction of a cytokine, induction of a chemokine, and expression of a cell surface marker.

49. A method comprising:

(a) identifying a peptide sequence associated with cancer and comprising a cancer mutation, wherein identifying the peptide sequence comprises comparing DNA, RNA, or protein sequences from cancer cells of the subject to DNA, RNA, or protein sequences from normal cells of the subject;

(b) selecting a protein encoded by an HLA class II allele that is normally expressed by a cell of the subject, and predicted by a machine-learning HLA-peptide presentation prediction model to bind to the peptide; wherein the prediction model has a positive predictive value of at least 0.1 at a recall rate of at least 0.1%, from 0.1%-50% or at most 50% and wherein the protein has a probability greater than a threshold presentation prediction probability value for presenting the identified peptide sequence;

(c) contacting the identified peptide with the selected protein encoded by the HLA class II allele to verify whether the identified peptide competes with a placeholder peptide associated with the selected protein encoded by the HLA class II allele to displace the placeholder peptide with an IC50 value of less than 500 nM; and

(d) optionally, purifying the identified peptide.

50. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:

(a) obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information associated with the HLA protein expressed in cells;

(c) determining or predicting that each of the plurality of peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the plurality of presentation predictions.

51. A method for manufacturing HLA class II tetramers by conjugation of four individual HLA protein alpha1 and beta1 heterodimers, the method comprising:

(a) expressing in a eukaryotic cell, a vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, wherein the heterodimer is associated with a placeholder peptide,

(b) purifying the secreted heterodimer from cell medium,

(c) validating the peptide binding activity using peptide exchange assay,

(d) adding streptavidin thereby conjugating heterodimers into tetramers,

(e) purifying the tetramers and having a yield of greater than 1 mg/L.

52. An HLA class II tetramer or multimer comprising either HLA-DR, or HLA-DP, or HLA-DQ heterodimers, each heterodimer comprising an alpha and a beta chain, wherein the heterodimer is purified and present at a concentration of greater than 1 mg/L.

53. The HLA class II tetramer or multimer of claim 52, wherein the HLA protein is an HLA class II protein selected from the group consisting of: HLA-DPB1*01:01/HLA-DPA1*01:03, HLA-DPB1*02:01/HLA-DPA1*01:03, HLA-DPB1*03:01/HLA-DPA1*01:03, HLA-DPB1*04:01/HLA-DPA1*01:03, HLA-DPB1*04:02/HLA-DPA1*01:03, HLA-DPB1*06:01/HLA-DPA1*01:03,HLA-DQB1*02:01/HLA-DQA1*05:01,HLA-DQB1*02:02/HLA-DQA1*02:01, HLA-DQB1*06:02/HLA-DQA1*01:02,HLA-DQB1*06:04/HLA-DQA1*01:02, HLA-DRB1*01:01, HLA-DRB1*01:02, HLA-DRB1*03:01, HLA-DRB1*03:02, HLA-DRB1*04:01, HLA-DRB1*04:02, HLA-DRB1*04:03, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*04:07, HLA-DRB1*07:01, HLA-DRB1*08:01, HLA-DRB1*08:02, HLA-DRB1*08:03, HLA-DRB1*08:04, HLA-DRB1*09:01, HLA-DRB1*10:01, HLA-DRB1*11:01, HLA-DRB1*11:02, HLA-DRB1*11:04, HLA-DRB1*12:01, HLA-DRB1*12:02, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*13:03, HLA-DRB1*14:01, HLA-DRB1*15:01, HLA-DRB1*15:02, HLA-DRB1*15:03, HLA-DRB1*16:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB3*03:01, HLA-DRB4*01:01, and HLA-DRB5*01:01.

54. A vector comprising a nucleic acid sequence encoding an alpha chain and a beta chain of HLA protein of any one of the preceding claims, a secretion signal, a biotinylation motif and at least one tag for identification or for purification, such that each HLA protein alpha 1 and beta1 heterodimers is secreted in dimerized state, optionally, wherein the secreted heterodimer is associated with a placeholder peptide.

55. A cell, comprising the vector of claim 54.

56. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, comprising:

(a) obtaining, by a computer processor, a plurality of peptide sequences of the polypeptide sequence;

(b) processing, by a computer processor, amino acid information of the plurality of peptide sequences using a machine-learning HLA-peptide presentation prediction model to generate a presentation prediction for each of the plurality of peptide sequences, each presentation prediction indicative of a likelihood that one or more proteins encoded by a class I or II MHC allele of a cell of the subject can present an epitope sequence of a given peptide sequence of the plurality of peptide sequences, wherein the machine-learning HLA-peptide presentation prediction model is trained using training data comprising sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; and

(c) determining or predicting that at least one of the plurality of peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the plurality of presentation predictions.

57. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:

(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data, wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry;

training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables; and

(b) determining or predicting that each of the peptide sequences of the polypeptide sequence would not be immunogenic to the subject based on the set of presentation predictions.

58. A method of screening a drug comprising a polypeptide sequence for immunogenicity in a subject, the method comprising:

(a) inputting amino acid information of peptide sequences of the polypeptide sequence, using a computer processor, into a machine-learning HLA-peptide presentation prediction model to generate a set of presentation predictions for the peptide sequences, each presentation prediction representing a probability that one or more proteins encoded by a class I or II MHC allele of a cell of the subject will present an epitope sequence of a given peptide sequence; wherein the machine-learning HLA-peptide presentation prediction model comprises: a plurality of predictor variables identified at least based on training data; wherein the training data comprises: sequence information of sequences of peptides presented by a HLA protein expressed in cells and identified by mass spectrometry; training peptide sequence information comprising amino acid position information, wherein the training peptide sequence information is associated with the HLA protein expressed in cells; and a function representing a relation between the amino acid position information received as input and the presentation likelihood generated as output based on the amino acid position information and the predictor variables;

(b) determining or predicting that at least one of the peptide sequences of the polypeptide sequence would be immunogenic to the subject based on the set of presentation predictions.

59. Use of a composition in the manufacture of a medicament for the treatment of an autoimmune disease or condition in a subject in need thereof, the composition comprising a regulatory T cell expressing a T cell receptor (TCR) that binds to a complex comprising: (i) an epitope of an expressed protein identified or predicted to be presented by a class I or II MHC of a cell of the subject, and (ii) the class I or II MHC, wherein the complex is targeted by a CD8 or CD4 T cell of the subject.