Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 3091917 A1 2019/09/06

(21) 3 091 917

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2019/02/27
(87) Date publication PCT/PCT Publication Date: 2019/09/06
(85) Entrée phase nationale/National Entry: 2020/08/20
(86) N° demande PCT/PCT Application No.: US 2019/019836
(87) N° publication PCT/PCT Publication No.: 2019/168984
(30) Priorité/Priority: 2018/02/27 (US62/636,061)

(51) Cl.Int./Int.Cl. *C07K 14/705* (2006.01), *C12Q 1/70* (2006.01), *C40B 20/06* (2006.01), *C40B 30/10* (2006.01), *G01N 27/49* (2006.01), *G01N 33/48* (2006.01)

(71) Demandeur/Applicant:
GRITSTONE ONCOLOGY, INC., US

(72) Inventeurs/Inventors:
BOUCHER, THOMAS FRANCIS, US;
BULIK-SULLIVAN, BRENDAN, US;
BUSBY, JENNIFER, US;
SKOBERNE, MOJCA, US;
YELENSKY, ROMAN, US

(74) Agent: GOWLING WLG (CANADA) LLP

(54) Titre : IDENTIFICATION DE NEO-ANTIGENE A L'AIDE DE MODELES PAN-ALLELES
(54) Title: NEOANTIGEN IDENTIFICATION WITH PAN-ALLELE MODELS

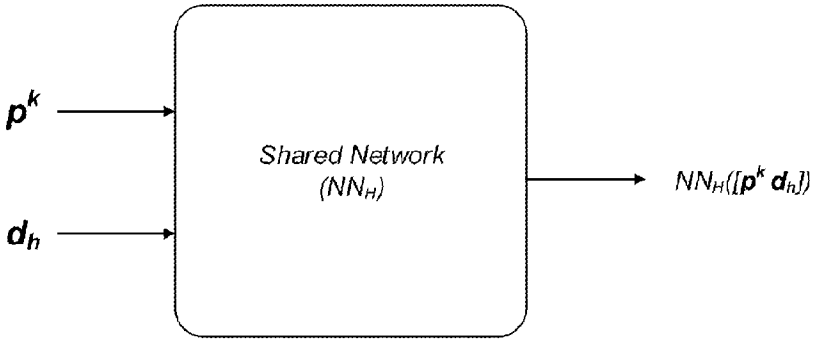


FIG. 13

(57) Abrégé/Abstract:
A method for identifying neoantigens that are likely to be presented by MHC alleles on a surface of tumor cells of a subject. Peptide sequences of the tumor neoantigens and of the MHC alleles are obtained by sequencing the tumor cells of the subject. The peptide sequences of the tumor neoantigens and of the MHC alleles are input into a machine-learned presentation model to generate presentation likelihoods for the tumor neoantigens, each presentation likelihood representing the likelihood that a neoantigen is presented by at least one of the MHC alleles on the surface of the tumor cells of the subject. A subset of the neoantigens is selected based on the presentation likelihoods.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
06 September 2019 (06.09.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/168984 A8

(51) International Patent Classification:
C07K 14/705 (2006.01) *C40B 30/10* (2006.01)
C12Q 1/70 (2006.01) *G01N 27/49* (2006.01)
C40B 20/06 (2006.01) *G01N 33/48* (2006.01)

(21) International Application Number:
PCT/US2019/019836

(22) International Filing Date:
27 February 2019 (27.02.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/636,061 27 February 2018 (27.02.2018) US

(71) Applicant: **GRITSTONE ONCOLOGY, INC.** [US/US]; 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US).

(72) Inventors: **BOUCHER, Thomas Francis**; Gritstone Oncology, Inc., 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US). **BULIK-SULLIVAN, Brendan**; Gritstone Oncology, Inc., 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US). **BUSBY, Jennifer**; Gritstone Oncology, Inc., 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US). **SKOBERNE, Mojca**; Gritstone Oncology, Inc., 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US). **YELENSKY, Roman**; Gritstone Oncology, Inc., 5959 Horton Street, Suite 300, Emeryville, CA 94608 (US).

(74) Agent: **KABLER, Kevin** et al.; Patent Administrator, GOODWIN PROCTER LLP, 100 Northern Avenue, Boston, Massachusetts 02210 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) Title: NEOANTIGEN IDENTIFICATION WITH PAN-ALLELE MODELS

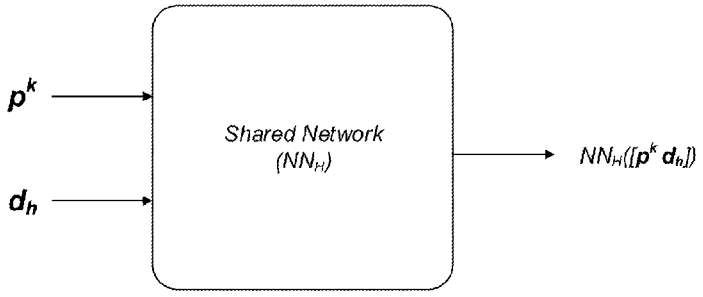


FIG. 13

(57) Abstract: A method for identifying neoantigens that are likely to be presented by MHC alleles on a surface of tumor cells of a subject. Peptide sequences of the tumor neoantigens and of the MHC alleles are obtained by sequencing the tumor cells of the subject. The peptide sequences of the tumor neoantigens and of the MHC alleles are input into a machine-learned presentation model to generate presentation likelihoods for the tumor neoantigens, each presentation likelihood representing the likelihood that a neoantigen is presented by at least one of the MHC alleles on the surface of the tumor cells of the subject. A subset of the neoantigens is selected based on the presentation likelihoods.

WO 2019/168984 A8

[Continued on next page]

WO 2019/168984 A8

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(48) Date of publication of this corrected version:

04 June 2020 (04.06.2020)

(15) Information about Correction:

see Notice of 04 June 2020 (04.06.2020)

DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 2
CONTENANT LES PAGES 1 À 165

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 2
CONTAINING PAGES 1 TO 165

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE:

NEOANTIGEN IDENTIFICATION WITH PAN-ALLELE MODELS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of and priority to U.S. Provisional Application No. 62/636,061, filed February 27, 2018. The content of the above referenced application is incorporated by reference in its entirety.

BACKGROUND

[0002] Therapeutic vaccines and T-cell therapy based on tumor-specific neoantigens hold great promise as a next-generation of personalized cancer immunotherapy. [1–3] Cancers with a high mutational burden, such as non-small cell lung cancer (NSCLC) and melanoma, are particularly attractive targets of such therapy given the relatively greater likelihood of neoantigen generation. [4,5] Early evidence shows that neoantigen-based vaccination can elicit T-cell responses[6] and that neoantigen targeted T-cell therapy can cause tumor regression under certain circumstances in selected patients.[7] Both MHC class I and MHC class II have an impact on T-cell responses[70-71].

[0003] However identification of neoantigens and neoantigen-recognizing T-cells has become a central challenge in assessing tumor responses[77,110], examining tumor evolution[111] and designing the next generation of personalized therapies[112]. Current neoantigen identification techniques are either time-consuming and laborious[84,96], or insufficiently precise[87,91–93]. Although it has recently been demonstrated that neoantigen-recognizing T-cells are a major component of TIL[84,96,113,114] and circulate in the peripheral blood of cancer patients[107], current methods for identifying neoantigen-reactive T-cells have some combination of the following three limitations: (1) they rely on difficult-to-obtain clinical specimens such as TIL[97,98] or leukaphereses[107] (2) they require screening impractically large libraries of peptides[95] or (3) they rely on MHC multimers, which may practically be available for only a small number of MHC alleles.

[0004] Furthermore, initial methods have been proposed incorporating mutation-based analysis using next-generation sequencing, RNA gene expression, and prediction of MHC binding affinity of candidate neoantigen peptides [8]. However, these proposed methods can fail to model the entirety of the epitope generation process, which contains many steps (e.g., TAP transport, proteasomal cleavage, MHC binding, transport of the peptide-MHC complex to the cell surface, and/or TCR recognition for MHC-I; endocytosis or autophagy, cleavage

via extracellular or lysosomal proteases (e.g., cathepsins), competition with the CLIP peptide for HLA-DM-catalyzed HLA binding, transport of the peptide-MHC complex to the cell surface and/or TCR recognition for MHC-II) in addition to gene expression and MHC binding[9]. Consequently, existing methods are likely to suffer from reduced low positive predictive value (PPV). (FIG. 1A)

[0005] Indeed, analyses of peptides presented by tumor cells performed by multiple groups have shown that <5% of peptides that are predicted to be presented using gene expression and MHC binding affinity can be found on the tumor surface MHC[10,11] (FIG. 1B). This low correlation between binding prediction and MHC presentation was further reinforced by recent observations of the lack of predictive accuracy improvement of binding-restricted neoantigens for checkpoint inhibitor response over the number of mutations alone.[12]

[0006] This low positive predictive value (PPV) of existing methods for predicting presentation presents a problem for neoantigen-based vaccine design and for neoantigen-based T-cell therapy. If vaccines are designed using predictions with a low PPV, most patients are unlikely to receive a therapeutic neoantigen and fewer still are likely to receive more than one (even assuming all presented peptides are immunogenic). Similarly, if therapeutic T-cells are designed based on predictions with a low PPV, most patients are unlikely to receive T-cells that are reactive to tumor neoantigens and the time and physical resource cost of identifying predictive neoantigens using downstream laboratory techniques post-prediction may be unduly high. Thus, neoantigen vaccination and T-cell therapy with current methods is unlikely to succeed in a substantial number of subjects having tumors. (FIG. 1C)

[0007] Additionally, previous approaches generated candidate neoantigens using only *cis*-acting mutations, and largely neglected to consider additional sources of neo-ORFs, including mutations in splicing factors, which occur in multiple tumor types and lead to aberrant splicing of many genes[13], and mutations that create or remove protease cleavage sites.

[0008] Finally, standard approaches to tumor genome and transcriptome analysis can miss somatic mutations that give rise to candidate neoantigens due to suboptimal conditions in library construction, exome and transcriptome capture, sequencing, or data analysis. Likewise, standard tumor analysis approaches can inadvertently promote sequence artifacts

or germline polymorphisms as neoantigens, leading to inefficient use of vaccine capacity or auto-immunity risk, respectively.

SUMMARY

[0009] Disclosed herein is an optimized approach for identifying and selecting neoantigens for personalized cancer vaccines, for T-cell therapy, or both. First, optimized tumor exome and transcriptome analysis approaches for neoantigen candidate identification using next-generation sequencing (NGS) are addressed. These methods build on standard approaches for NGS tumor analysis to ensure that the highest sensitivity and specificity neoantigen candidates are advanced, across all classes of genomic alteration. Second, novel approaches for high-PPV neoantigen selection are presented to overcome the specificity problem and ensure that neoantigens advanced for vaccine inclusion and/or as targets for T-cell therapy are more likely to elicit anti-tumor immunity. These approaches include, depending on the embodiment, a trained statistical regression or nonlinear deep learning model that is configured to predict presentation of peptides of multiple lengths, sharing statistical strength across peptides of different lengths, on a pan-allele basis. The model is capable of predicting the probability that a peptide will be presented by *any* MHC allele—including unknown MHC alleles that the model has not previously encountered during training. The nonlinear deep learning models particularly can be designed and trained to treat different MHC alleles in the same cell as independent, thereby addressing problems with linear models that would have them interfere with each other. Finally, additional considerations for personalized vaccine design and manufacturing based on neoantigens, and for production of personalized neoantigen-specific T-cells for T-cell therapy, are addressed.

[0010] The model disclosed herein outperforms state-of-the-art predictors trained on binding affinity and early predictors based on MS peptide data by up to an order of magnitude. By more reliably predicting presentation of peptides, the model enables more time- and cost-effective identification of neoantigen-specific or tumor antigen-specific T-cells for personalized therapy using a clinically practical process that uses limited volumes of patient peripheral blood, screens few peptides per patient, and does not necessarily rely on MHC multimers. However, in another embodiment, the model disclosed herein can be used to enable more time- and cost-effective identification of tumor antigen-specific T cells using MHC multimers, by decreasing the number of peptides bound to MHC multimers that need to be screened in order to identify neoantigen- or tumor antigen-specific T cells.

[0011] The predictive performance of the model disclosed herein on the TIL neoepitope dataset and the prospective neoantigen-reactive T-cell identification task demonstrate that it is now possible to obtain therapeutically-useful neoepitope predictions by modeling HLA processing and presentation. In summary, this work offers practical *in silico* antigen identification for antigen-targeted immunotherapy, thereby accelerating progress towards cures for patients.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0012] These and other features, aspects, and advantages of the present invention will become better understood with regard to the following description, and accompanying drawings, where:

[0013] FIG. 1A shows current clinical approaches to neoantigen identification.

[0014] FIG. 1B shows that <5% of predicted bound peptides are presented on tumor cells.

[0015] FIG. 1C shows the impact of the neoantigen prediction specificity problem.

[0016] FIG. 1D shows that binding prediction is not sufficient for neoantigen identification.

[0017] FIG. 1E shows probability of MHC-I presentation as a function of peptide length.

[0018] FIG. 1F shows an example peptide spectrum generated from Promega's dynamic range standard.

[0019] FIG. 1G shows how the addition of features increases the model positive predictive value.

[0020] FIG. 2A is an overview of an environment for identifying likelihoods of peptide presentation in patients, in accordance with an embodiment.

[0021] FIGS. 2B and 2C illustrate a method of obtaining presentation information, in accordance with an embodiment.

[0022] FIG. 3 is a high-level block diagram illustrating the computer logic components of the presentation identification system, according to one embodiment.

[0023] FIG. 4 illustrates an example set of training data, according to one embodiment.

[0024] FIG. 5 illustrates an example network model in association with an MHC allele.

[0025] FIG. 6 illustrates an example network model $NN_H(\cdot)$ shared by MHC alleles, according to one embodiment.

[0026] FIG. 7 illustrates generating a presentation likelihood for a peptide in association with an MHC allele using an example network model.

[0027] FIG. 8 illustrates generating a presentation likelihood for a peptide in association with a MHC allele using example network models.

[0028] FIG. 9 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.

[0029] FIG. 10 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.

[0030] FIG. 11 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.

[0031] FIG. 12 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.

[0032] FIG. 13 illustrates an example network model $NN_H(\cdot)$ shared by MHC alleles, according to an embodiment.

[0033] FIG. 14 illustrates an example network model that is not associated with an MHC allele.

[0034] FIG. 15 illustrates generating a presentation likelihood for a peptide in association with an MHC allele using an example network model shared by MHC alleles.

[0035] FIG. 16 illustrates a precision/recall curve output by a pan-allele model comprising a neural network and trained on samples that include a tested HLA allele, and a precision/recall curve output by a pan-allele model comprising a neural network and not trained on samples that include the tested HLA allele, for a first test sample.

[0036] FIG. 17 illustrates a precision/recall curve output by a pan-allele model comprising a neural network and trained on samples that include a tested HLA allele, and a precision/recall curve output by a pan-allele model comprising a neural network and not trained on samples that include the tested HLA allele, for a second test sample.

[0037] FIG. 18 illustrates a precision/recall curve output by a pan-allele model comprising a neural network and trained on samples that include a tested HLA allele, and a precision/recall curve output by a pan-allele model comprising a neural network and not trained on samples that include the tested HLA allele, for a third test sample.

[0038] FIG. 19 illustrates precision/recall curves output by a pan-allele model comprising a neural network, a random forest model, a quadratic discriminant model, and a MHCFlurry model trained on samples that include a tested HLA allele.

[0039] FIG. 20 illustrates precision/recall curves output by a pan-allele model comprising a neural network, a random forest model, a quadratic discriminant model, and a MHCFlurry model not trained on samples that include a tested HLA allele, for a first test sample.

[0040] FIG. 21 illustrates precision/recall curves output by a pan-allele model comprising a neural network, a random forest model, a quadratic discriminant model, and a MHCFlurry model not trained on samples that include a tested HLA allele, for a second test sample.

[0041] FIG. 22 illustrates precision/recall curves output by a pan-allele model comprising a neural network, a random forest model, a quadratic discriminant model, and a MHCFlurry model not trained on samples that include a tested HLA allele, for a third test sample.

[0042] FIG. 23A illustrates a sample frequency distribution of mutation burden in NSCLC patients.

[0043] FIG. 23B illustrates the number of presented neoantigens in simulated vaccines for patients selected based on an inclusion criteria of whether the patients satisfy a minimum mutation burden, in accordance with an embodiment.

[0044] FIG. 23C compares the number of presented neoantigens in simulated vaccines between selected patients associated with vaccines including treatment subsets identified based on presentation models and selected patients associated with vaccines including treatment subsets identified through current state-of-the-art models, in accordance with an embodiment.

[0045] FIG. 23D compares the number of presented neoantigens in simulated vaccines between selected patients associated with vaccines including treatment subsets identified based on a single per-allele presentation model for HLA-A*02:01 and selected patients associated with vaccines including treatment subsets identified based on both per-allele presentation models for HLA-A*02:01 and HLA-B*07:02. The vaccine capacity is set as v=20 epitopes, in accordance with an embodiment.

[0046] FIG. 23E compares the number of presented neoantigens in simulated vaccines between patients selected based on mutation burden and patients selected by expectation utility score, in accordance with an embodiment.

[0047] FIG. 24 compares the positive predictive values (PPVs) at 40% recall of a pan-allele presentation model that uses presentation hotspot parameters and a pan-allele presentation model that does not use presentation hotspot parameters, when the pan-allele models are tested on five held-out test samples.

[0048] FIG. 25A compares the proportion of somatic mutations recognized by T-cells (e.g., pre-existing T-cell responses) for the top 5, 10, and 20-ranked somatic mutations identified using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the allele-specific neural network model, and the pan-allele neural network model for a test set comprising 12 different test samples, each test sample taken from a patient with at least one pre-existing T-cell response.

[0049] FIG. 25B compares the proportion of minimal neoepitopes recognized by T-cells (e.g., pre-existing T-cell responses) for the top 5, 10, and 20-ranked minimal neoepitopes identified using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the allele-specific neural network model, and the pan-allele neural network model for a test set comprising 12 different test samples, each test sample taken from a patient with at least one pre-existing T-cell response.

[0050] FIG. 26A depicts detection of T-cell responses to patient-specific neoantigen peptide pools for nine patients.

[0051] FIG. 26B depicts detection of T-cell responses to individual patient-specific neoantigen peptides for four patients.

[0052] FIG. 26C depicts example images of ELISpot wells for patient CU04.

[0053] FIG. 27A depicts results from control experiments with neoantigens in HLA-matched healthy donors.

[0054] FIG. 27B depicts results from control experiments with neoantigens in HLA-matched healthy donors.

[0055] FIG. 28 depicts detection of T-cell responses to PHA positive control for each donor and each *in vitro* expansion depicted in FIG. 26A.

[0056] FIG. 29A depicts detection of T-cell responses to each individual patient-specific neoantigen peptide in pool #2 for patient CU04.

[0057] FIG. 29B depicts detection of T-cell responses to individual patient-specific neoantigen peptides for each of three visits of patient CU04 and for each of two visits of patient 1-024-002, each visit occurring at a different time point.

[0058] FIG. 29C depicts detection of T-cell responses to individual patient-specific neoantigen peptides and to patient-specific neoagntigen peptide pools for each of two visits of patient CU04 and for each of two visits of patient 1-024-002, each visit occurring at a different time point.

[0059] FIG. 30 depicts detection of T-cell responses to the two patient-specific neoantigen peptide pools and to DMSO negative controls for the patients of FIG. 26A.

[0060] FIG. 31A depicts the precision-recall curves for each of the test sample 0 including class II MHC alleles for the pan-allele and the allele-specific models.

[0061] FIG. 31B depicts the precision-recall curves for each of the test sample 1 including class II MHC alleles for the pan-allele and the allele-specific models.

[0062] FIG. 31C depicts the precision-recall curves for each of the test sample 2 including class II MHC alleles for the pan-allele and the allele-specific models.

[0063] FIG. 31D depicts the precision-recall curves for each of the test sample 4 including class II MHC alleles for the pan-allele and the allele-specific models.

[0064] FIG. 32 depicts a method for sequencing TCRs of neoantigen-specific memory T-cells from the peripheral blood of a NSCLC patient.

[0065] FIG. 33 depicts exemplary embodiments of TCR constructs for introducing a TCR into recipient cells.

[0066] FIG. 34 depicts an exemplary P526 construct backbone nucleotide sequence for cloning TCRs into expression systems for therapy development.

[0067] FIG. 35 depicts an exemplary construct sequence for cloning patient neoantigen-specific TCR, clonotype 1 TCR into expression systems for therapy development.

[0068] FIG. 36 depicts an exemplary construct sequence for cloning patient neoantigen-specific TCR, clonotype 3 into expression systems for therapy development.

[0069] FIG. 37 is a flow chart of a method for providing a customized, neoantigen-specific treatment to a patient, in accordance with an embodiment.

[0070] FIG. 38 illustrates an example computer for implementing the entities shown in FIGS. 1 and 3.

DETAILED DESCRIPTION

I. Definitions

[0071] In general, terms used in the claims and the specification are intended to be construed as having the plain meaning understood by a person of ordinary skill in the art. Certain terms are defined below to provide additional clarity. In case of conflict between the plain meaning and the provided definitions, the provided definitions are to be used.

[0072] As used herein the term "antigen" is a substance that induces an immune response.

[0073] As used herein the term "neoantigen" is an antigen that has at least one alteration that makes it distinct from the corresponding wild-type, parental antigen, e.g., via mutation in a tumor cell or post-translational modification specific to a tumor cell. A neoantigen can include a polypeptide sequence or a nucleotide sequence. A mutation can include a frameshift or nonframeshift indel, missense or nonsense substitution, splice site alteration, genomic rearrangement or gene fusion, or any genomic or expression alteration giving rise to a neoORF. A mutations can also include a splice variant. Post-translational modifications specific to a tumor cell can include aberrant phosphorylation. Post-translational modifications specific to a tumor cell can also include a proteasome-generated spliced antigen. *See* Liepe et al., A large fraction of HLA class I ligands are proteasome-generated spliced peptides; Science. 2016 Oct 21;354(6310):354-358.

[0074] As used herein the term "tumor neoantigen" is a neoantigen present in a subject's tumor cell or tissue but not in the subject's corresponding normal cell or tissue.

[0075] As used herein the term "neoantigen-based vaccine" is a vaccine construct based on one or more neoantigens, e.g., a plurality of neoantigens.

[0076] As used herein the term "candidate neoantigen" is a mutation or other aberration giving rise to a new sequence that may represent a neoantigen.

[0077] As used herein the term "coding region" is the portion(s) of a gene that encode protein.

[0078] As used herein the term "coding mutation" is a mutation occurring in a coding region.

[0079] As used herein the term "ORF" means open reading frame.

[0080] As used herein the term "NEO-ORF" is a tumor-specific ORF arising from a mutation or other aberration such as splicing.

[0081] As used herein the term "missense mutation" is a mutation causing a substitution from one amino acid to another.

[0082] As used herein the term "nonsense mutation" is a mutation causing a substitution from an amino acid to a stop codon.

[0083] As used herein the term "frameshift mutation" is a mutation causing a change in the frame of the protein.

[0084] As used herein the term "indel" is an insertion or deletion of one or more nucleic acids.

[0085] As used herein, the term percent "identity," in the context of two or more nucleic acid or polypeptide sequences, refer to two or more sequences or subsequences that have a specified percentage of nucleotides or amino acid residues that are the same, when compared and aligned for maximum correspondence, as measured using one of the sequence comparison algorithms described below (e.g., BLASTP and BLASTN or other algorithms available to persons of skill) or by visual inspection. Depending on the application, the percent "identity" can exist over a region of the sequence being compared, e.g., over a functional domain, or, alternatively, exist over the full length of the two sequences to be compared.

[0086] For sequence comparison, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters. Alternatively, sequence similarity or dissimilarity can be established by the combined presence or absence of particular nucleotides, or, for translated sequences, amino acids at selected sequence positions (e.g., sequence motifs).

[0087] Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol. 48:443 (1970), by the search for similarity method of Pearson & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, Wis.), or by visual inspection (see generally Ausubel et al., infra).

[0088] One example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul et al., J. Mol. Biol. 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information.

[0089] As used herein the term “non-stop or read-through” is a mutation causing the removal of the natural stop codon.

[0090] As used herein the term “epitope” is the specific portion of an antigen typically bound by an antibody or T-cell receptor.

[0091] As used herein the term "immunogenic" is the ability to elicit an immune response, e.g., via T-cells, B cells, or both.

[0092] As used herein the term "HLA binding affinity" "MHC binding affinity" means affinity of binding between a specific antigen and a specific MHC allele.

[0093] As used herein the term "bait" is a nucleic acid probe used to enrich a specific sequence of DNA or RNA from a sample.

[0094] As used herein the term "variant" is a difference between a subject's nucleic acids and the reference human genome used as a control.

[0095] As used herein the term "variant call" is an algorithmic determination of the presence of a variant, typically from sequencing.

[0096] As used herein the term "polymorphism" is a germline variant, i.e., a variant found in all DNA-bearing cells of an individual.

[0097] As used herein the term "somatic variant" is a variant arising in non-germline cells of an individual.

[0098] As used herein the term "allele" is a version of a gene or a version of a genetic sequence or a version of a protein.

[0099] As used herein the term "HLA type" is the complement of HLA gene alleles.

[00100] As used herein the term "nonsense-mediated decay" or "NMD" is a degradation of an mRNA by a cell due to a premature stop codon.

[00101] As used herein the term "truncal mutation" is a mutation originating early in the development of a tumor and present in a substantial portion of the tumor's cells.

[00102] As used herein the term "subclonal mutation" is a mutation originating later in the development of a tumor and present in only a subset of the tumor's cells.

[00103] As used herein the term "exome" is a subset of the genome that codes for proteins. An exome can be the collective exons of a genome.

[00104] As used herein the term "logistic regression" is a regression model for binary data from statistics where the logit of the probability that the dependent variable is equal to one is modeled as a linear function of the dependent variables.

[00105] As used herein the term "neural network" is a machine learning model for classification or regression consisting of multiple layers of linear transformations followed by element-wise nonlinearities typically trained via stochastic gradient descent and back-propagation.

[00106] As used herein the term "proteome" is the set of all proteins expressed and/or translated by a cell, group of cells, or individual.

[00107] As used herein the term "peptidome" is the set of all peptides presented by MHC-I or MHC-II on the cell surface. The peptidome may refer to a property of a cell or a collection of cells (e.g., the tumor peptidome, meaning the union of the peptidomes of all cells that comprise the tumor).

[00108] As used herein the term "ELISPOT" means Enzyme-linked immunosorbent spot assay – which is a common method for monitoring immune responses in humans and animals.

[00109] As used herein the term "dextramers" is a dextran-based peptide-MHC multimers used for antigen-specific T-cell staining in flow cytometry.

[00110] As used herein the term "MHC multimers" is a peptide-MHC complex comprising multiple peptide- MHC monomer units.

[00111] As used herein the term "MHC tetramers" is a peptide-MHC complex comprising four peptide- MHC monomer units.

[00112] As used herein the term "tolerance or immune tolerance" is a state of immune non-responsiveness to one or more antigens, e.g. self-antigens.

[00113] As used herein the term "central tolerance" is a tolerance affected in the thymus, either by deleting self-reactive T-cell clones or by promoting self-reactive T-cell clones to differentiate into immunosuppressive regulatory T-cells (Tregs).

[00114] As used herein the term "peripheral tolerance" is a tolerance affected in the periphery by downregulating or anergizing self-reactive T-cells that survive central tolerance or promoting these T-cells to differentiate into Tregs.

[00115] The term "sample" can include a single cell or multiple cells or fragments of cells or an aliquot of body fluid, taken from a subject, by means including venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage sample, scraping, surgical incision, or intervention or other means known in the art.

[00116] The term "subject" encompasses a cell, tissue, or organism, human or non-human, whether in vivo, ex vivo, or in vitro, male or female. The term subject is inclusive of mammals including humans.

[00117] The term "mammal" encompasses both humans and non-humans and includes but is not limited to humans, non-human primates, canines, felines, murines, bovines, equines, and porcines.

[00118] The term "clinical factor" refers to a measure of a condition of a subject, e.g., disease activity or severity. "Clinical factor" encompasses all markers of a subject's health status, including non-sample markers, and/or other characteristics of a subject, such as, without limitation, age and gender. A clinical factor can be a score, a value, or a set of values that can be obtained from evaluation of a sample (or population of samples) from a subject or a subject under a determined condition. A clinical factor can also be predicted by markers and/or other parameters such as gene expression surrogates. Clinical factors can include tumor type, tumor sub-type, and smoking history.

[00119] Abbreviations: MHC: major histocompatibility complex; HLA: human leukocyte antigen, or the human MHC gene locus; NGS: next-generation sequencing; PPV: positive predictive value; TSNA: tumor-specific neoantigen; FFPE: formalin-fixed, paraffin-embedded; NMD: nonsense-mediated decay; NSCLC: non-small-cell lung cancer; DC: dendritic cell.

[00120] It should be noted that, as used in the specification and the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise.

[00121] Any terms not directly defined herein shall be understood to have the meanings commonly associated with them as understood within the art of the invention. Certain terms are discussed herein to provide additional guidance to the practitioner in describing the compositions, devices, methods and the like of aspects of the invention, and how to make or use them. It will be appreciated that the same thing may be said in more than one way. Consequently, alternative language and synonyms may be used for any one or more of the terms discussed herein. No significance is to be placed upon whether or not a term is elaborated or discussed herein. Some synonyms or substitutable methods, materials and the like are provided. Recital of one or a few synonyms or equivalents does not exclude use of other synonyms or equivalents, unless it is explicitly stated. Use of examples, including examples of terms, is for illustrative purposes only and does not limit the scope and meaning of the aspects of the invention herein.

[00122] All references, issued patents and patent applications cited within the body of the specification are hereby incorporated by reference in their entirety, for all purposes.

II. Methods of Identifying Neoantigens

[00123] Disclosed herein are methods for identifying at least one neoantigen from one or more tumor cells of a subject that are likely to be presented by one or more MHC alleles on a

surface of the tumor cells. The method includes obtaining exome, transcriptome, and/or whole genome nucleotide sequencing data from the tumor cells as well as normal cells of the subject. This nucleotide sequencing data is used to obtain a peptide sequence of each neoantigen in a set of neoantigens. The set of neoantigens is identified by comparing the nucleotide sequencing data from the tumor cells and the nucleotide sequencing data from the normal cells. Specifically, the peptide sequence of each neoantigen in the set of neoantigens comprises at least one alteration that makes it distinct from the corresponding wild-type peptide sequence identified from the normal cells of the subject. The method further includes encoding the peptide sequence of each neoantigen in the set of neoantigens into a corresponding numerical vector. Each numerical vector includes information describing the amino acids that make up the peptide sequence and the positions of the amino acids in the peptide sequence. The method further comprises obtaining exome, transcriptome, and/or whole genome nucleotide sequencing data from the tumor cells of the subject. This nucleotide sequencing data is used to obtain a peptide sequence of each of the one or more MHC alleles of the subject. The peptide sequence of each of the one or more MHC alleles of the subject is encoded into a corresponding numerical vector. Each numerical vector includes information describing the amino acids that make up the peptide sequence of the MHC allele and the positions of the amino acids in the peptide sequence of the MHC allele. The method further comprises inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into a machine-learned presentation model to generate a presentation likelihood for each neoantigen in the set of neoantigens. Each presentation likelihood represents the likelihood that the corresponding neoantigen is presented by the one or more MHC alleles on the surface of the tumor cells of the subject. The machine-learned presentation model comprises a plurality of parameters and a function. The plurality of parameters are identified based on a training data set. The training data set comprises, for each sample in a plurality of samples, a label obtained by mass spectrometry measuring presence of peptides bound to at least one MHC allele in a set of MHC alleles identified as present in the sample, training peptide sequences encoded as numerical vectors that include information describing the amino acids that make up the peptides and the positions of the amino acids in the peptides, and training peptide sequences encoded as numerical vectors that include information describing the amino acids that make up the at least one MHC allele bound to the peptides of the sample and the positions of the amino acids in MHC allele

peptides. The function represents a relation between the numerical vectors received as input by the machine-learned presentation model and the presentation likelihood generated as output by the machine-learned presentation model based on the numerical vectors and the plurality of parameters. The method further includes selecting a subset of the set of neoantigens, based on the presentation likelihoods, to generate a set of selected neoantigens, and returning the set of selected neoantigens.

[00124] In some embodiments, inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model comprises applying the machine-learned presentation model to the peptide sequence of the neoantigen and to the peptide sequence of the one or more MHC alleles to generate a dependency score for each of the one or more MHC alleles. The dependency score for an MHC allele indicates whether the MHC allele will present the neoantigen, based on the particular amino acids at the particular positions of the peptide sequence. In further embodiments, inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model further comprises transforming the dependency scores to generate a corresponding per-allele likelihood for each MHC allele indicating a likelihood that the corresponding MHC allele will present the corresponding neoantigen, and combining the per-allele likelihoods to generate the presentation likelihood of the neoantigen. In some embodiments, transforming the dependency scores models the presentation of the neoantigen as mutually exclusive across the one or more MHC alleles. In alternative embodiments, inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model further comprises transforming a combination of the dependency scores to generate the presentation likelihood. In such embodiments, transforming the combination of the dependency scores models the presentation of the neoantigen as interfering between the one or more MHC alleles.

[00125] In some embodiments, the set of presentation likelihoods are further identified by one or more allele noninteracting features. In such embodiments, the method further comprises applying the machine-learned presentation model to the allele noninteracting features to generate a dependency score for the allele noninteracting features. The

dependency score indicates whether the peptide sequence of the corresponding neoantigen will be presented based on the allele noninteracting features. In some embodiments, the method further comprises combining the dependency score for each MHC allele of the one or more MHC alleles with the dependency score for the allele noninteracting features, transforming the combined dependency score for each MHC allele to generate a per-allele likelihood for each MHC allele, and combining the per-allele likelihoods to generate the presentation likelihood. The per-allele likelihood for a MHC allele indicates a likelihood that the MHC allele will present the corresponding neoantigen. In alternative embodiments, the method further comprises combining the dependency scores for each of the MHC alleles and the dependency score for the allele noninteracting features, and transforming the combined dependency scores to generate the presentation likelihood.

[00126] In some embodiments, the one or more MHC alleles include two or more different MHC alleles.

[00127] In some embodiments, the peptide sequences comprise peptide sequences having lengths other than 9 amino acids.

[00128] In some embodiments, encoding a peptide sequence comprises encoding the peptide sequence using a one-hot encoding scheme.

[00129] In some embodiments, the plurality of samples comprise at least one of cell lines engineered to express a single MHC allele, cell lines engineered to express a plurality of MHC alleles, human cell lines obtained or derived from a plurality of patients, fresh or frozen tumor samples obtained from a plurality of patients, and fresh or frozen tissue samples obtained from a plurality of patients.

[00130] In some embodiments, the training data set further comprises at least one of data associated with peptide-MHC binding affinity measurements for at least one of the peptides, and data associated with peptide-MHC binding stability measurements for at least one of the peptides.

[00131] In some embodiments, the set of presentation likelihoods are further identified by expression levels of the one or more MHC alleles in the subject, as measured by RNA-seq or mass spectrometry.

[00132] In some embodiments, the set of presentation likelihoods are further identified by features comprising at least one of predicted affinity between a neoantigen in the set of neoantigens and the one or more MHC alleles, and predicted stability of the neoantigen encoded peptide-MHC complex.

[00133] In some embodiments, the set of numerical likelihoods are further identified by features comprising at least one of the C-terminal sequences flanking the neoantigen encoded peptide sequence within its source protein sequence, and the N-terminal sequences flanking the neoantigen encoded peptide sequence within its source protein sequence.

[00134] In some embodiments, selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being presented on the tumor cell surface relative to unselected neoantigens, based on the machine-learned presentation model.

[00135] In some embodiments, selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being capable of inducing a tumor-specific immune response in the subject relative to unselected neoantigens, based on the machine-learned presentation model.

[00136] In some embodiments, selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being capable of being presented to naïve T-cells by professional antigen presenting cells (APCs) relative to unselected neoantigens, based on the presentation model. In such embodiments, the APC is optionally a dendritic cell (DC).

[00137] In some embodiments, selecting the set of selected neoantigens comprises selecting neoantigens that have a decreased likelihood of being subject to inhibition via central or peripheral tolerance relative to unselected neoantigens, based on the machine-learned presentation model.

[00138] In some embodiments, selecting the set of selected neoantigens comprises selecting neoantigens that have a decreased likelihood of being capable of inducing an autoimmune response to normal tissue in the subject relative to unselected neoantigens, based on the machine-learned presentation model.

[00139] In some embodiments, the one or more tumor cells are selected from the group consisting of: lung cancer, melanoma, breast cancer, ovarian cancer, prostate cancer, kidney cancer, gastric cancer, colon cancer, testicular cancer, head and neck cancer, pancreatic cancer, brain cancer, B-cell lymphoma, acute myelogenous leukemia, chronic myelogenous leukemia, chronic lymphocytic leukemia, and T-cell lymphocytic leukemia, non-small cell lung cancer, and small cell lung cancer.

[00140] In some embodiments, the method further comprises generating an output for constructing a personalized cancer vaccine from the set of selected neoantigens. In such embodiments, the output for the personalized cancer vaccine may comprise at least one

peptide sequence or at least one nucleotide sequence encoding the set of selected neoantigens.

[00141] In some embodiments, the machine-learned presentation model is a neural network model. In such embodiments, the neural network model may be a single neural network model that includes a series of nodes arranged in one or more layers. The single neural network model may be configured to receive numerical vectors encoding the peptide sequences of multiple different MHC alleles. In such embodiments, the neural network model may be trained by updating the parameters of the neural network model. In some embodiments, the machine-learned presentation model may be a deep learning model that includes one or more layers of nodes.

[00142] In some embodiments, the training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the at least one MHC allele bound to the peptides of the sample and a set of positions of the amino acids in the at least one MHC allele, do not include a peptide sequence of a MHC allele of the subject that is input into the machine-learned presentation model to generate the set of presentation likelihoods for the set of neoantigens.

[00143] In certain aspects disclosed herein, the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set belongs to a gene family to which the one or more MHC alleles of the subject belongs.

[00144] In some embodiments, the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set comprises one MHC allele. In alternative embodiments, the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set comprises more than one MHC allele.

[00145] In some embodiments, the one or more MHC alleles are class I MHC alleles.

[00146] Disclosed herein are also computer systems comprising a computer processor and a memory that stores computer program instructions that when executed by the computer processor, cause the computer processor to execute an embodiment of the method described above.

III. Identification of Tumor Specific Mutations in Neoantigens

[00147] Also disclosed herein are methods for the identification of certain mutations (e.g., the variants or alleles that are present in cancer cells). In particular, these mutations can be present in the genome, transcriptome, proteome, or exome of cancer cells of a subject having cancer but not in normal tissue from the subject.

[00148] Genetic mutations in tumors can be considered useful for the immunological targeting of tumors if they lead to changes in the amino acid sequence of a protein exclusively in the tumor. Useful mutations include: (1) non-synonymous mutations leading to different amino acids in the protein; (2) read-through mutations in which a stop codon is modified or deleted, leading to translation of a longer protein with a novel tumor-specific sequence at the C-terminus; (3) splice site mutations that lead to the inclusion of an intron in the mature mRNA and thus a unique tumor-specific protein sequence; (4) chromosomal rearrangements that give rise to a chimeric protein with tumor-specific sequences at the junction of 2 proteins (i.e., gene fusion); (5) frameshift mutations or deletions that lead to a new open reading frame with a novel tumor-specific protein sequence. Mutations can also include one or more of nonframeshift indel, missense or nonsense substitution, splice site alteration, genomic rearrangement or gene fusion, or any genomic or expression alteration giving rise to a neoORF.

[00149] Peptides with mutations or mutated polypeptides arising from for example, splice-site, frameshift, readthrough, or gene fusion mutations in tumor cells can be identified by sequencing DNA, RNA or protein in tumor versus normal cells.

[00150] Also mutations can include previously identified tumor specific mutations. Known tumor mutations can be found at the Catalogue of Somatic Mutations in Cancer (COSMIC) database.

[00151] A variety of methods are available for detecting the presence of a particular mutation or allele in an individual's DNA or RNA. Advancements in this field have provided accurate, easy, and inexpensive large-scale SNP genotyping. For example, several techniques have been described including dynamic allele-specific hybridization (DASH), microplate array diagonal gel electrophoresis (MADGE), pyrosequencing, oligonucleotide-specific ligation, the TaqMan system as well as various DNA "chip" technologies such as the Affymetrix SNP chips. These methods utilize amplification of a target genetic region, typically by PCR. Still other methods, based on the generation of small signal molecules by invasive cleavage followed by mass spectrometry or immobilized padlock probes and rolling-circle amplification. Several of the methods known in the art for detecting specific mutations are summarized below.

[00152] PCR based detection means can include multiplex amplification of a plurality of markers simultaneously. For example, it is well known in the art to select PCR primers to generate PCR products that do not overlap in size and can be analyzed simultaneously.

Alternatively, it is possible to amplify different markers with primers that are differentially labeled and thus can each be differentially detected. Of course, hybridization based detection means allow the differential detection of multiple PCR products in a sample. Other techniques are known in the art to allow multiplex analyses of a plurality of markers.

[00153] Several methods have been developed to facilitate analysis of single nucleotide polymorphisms in genomic DNA or cellular RNA. For example, a single base polymorphism can be detected by using a specialized exonuclease-resistant nucleotide, as disclosed, e.g., in Mundy, C. R. (U.S. Pat. No. 4,656,127). According to the method, a primer complementary to the allelic sequence immediately 3' to the polymorphic site is permitted to hybridize to a target molecule obtained from a particular animal or human. If the polymorphic site on the target molecule contains a nucleotide that is complementary to the particular exonuclease-resistant nucleotide derivative present, then that derivative will be incorporated onto the end of the hybridized primer. Such incorporation renders the primer resistant to exonuclease, and thereby permits its detection. Since the identity of the exonuclease-resistant derivative of the sample is known, a finding that the primer has become resistant to exonucleases reveals that the nucleotide(s) present in the polymorphic site of the target molecule is complementary to that of the nucleotide derivative used in the reaction. This method has the advantage that it does not require the determination of large amounts of extraneous sequence data.

[00154] A solution-based method can be used for determining the identity of a nucleotide of a polymorphic site. Cohen, D. et al. (French Patent 2,650,840; PCT Appln. No. WO91/02087). As in the Mundy method of U.S. Pat. No. 4,656,127, a primer is employed that is complementary to allelic sequences immediately 3' to a polymorphic site. The method determines the identity of the nucleotide of that site using labeled dideoxynucleotide derivatives, which, if complementary to the nucleotide of the polymorphic site will become incorporated onto the terminus of the primer. An alternative method, known as Genetic Bit Analysis or GBA is described by Goelet, P. et al. (PCT Appln. No. 92/15712). The method of Goelet, P. et al. uses mixtures of labeled terminators and a primer that is complementary to the sequence 3' to a polymorphic site. The labeled terminator that is incorporated is thus determined by, and complementary to, the nucleotide present in the polymorphic site of the target molecule being evaluated. In contrast to the method of Cohen et al. (French Patent 2,650,840; PCT Appln. No. WO91/02087) the method of Goelet, P. et al. can be a heterogeneous phase assay, in which the primer or the target molecule is immobilized to a solid phase.

[00155] Several primer-guided nucleotide incorporation procedures for assaying polymorphic sites in DNA have been described (Komher, J. S. et al., Nucl. Acids. Res. 17:7779-7784 (1989); Sokolov, B. P., Nucl. Acids Res. 18:3671 (1990); Syvanen, A.-C., et al., Genomics 8:684-692 (1990); Kuppuswamy, M. N. et al., Proc. Natl. Acad. Sci. (U.S.A.) 88:1143-1147 (1991); Prezant, T. R. et al., Hum. Mutat. 1:159-164 (1992); Ugozzoli, L. et al., GATA 9:107-112 (1992); Nyren, P. et al., Anal. Biochem. 208:171-175 (1993)). These methods differ from GBA in that they utilize incorporation of labeled deoxynucleotides to discriminate between bases at a polymorphic site. In such a format, since the signal is proportional to the number of deoxynucleotides incorporated, polymorphisms that occur in runs of the same nucleotide can result in signals that are proportional to the length of the run (Syvanen, A.-C., et al., Amer. J. Hum. Genet. 52:46-59 (1993)).

[00156] A number of initiatives obtain sequence information directly from millions of individual molecules of DNA or RNA in parallel. Real-time single molecule sequencing-by-synthesis technologies rely on the detection of fluorescent nucleotides as they are incorporated into a nascent strand of DNA that is complementary to the template being sequenced. In one method, oligonucleotides 30-50 bases in length are covalently anchored at the 5' end to glass cover slips. These anchored strands perform two functions. First, they act as capture sites for the target template strands if the templates are configured with capture tails complementary to the surface-bound oligonucleotides. They also act as primers for the template directed primer extension that forms the basis of the sequence reading. The capture primers function as a fixed position site for sequence determination using multiple cycles of synthesis, detection, and chemical cleavage of the dye-linker to remove the dye. Each cycle consists of adding the polymerase/labeled nucleotide mixture, rinsing, imaging and cleavage of dye. In an alternative method, polymerase is modified with a fluorescent donor molecule and immobilized on a glass slide, while each nucleotide is color-coded with an acceptor fluorescent moiety attached to a gamma-phosphate. The system detects the interaction between a fluorescently-tagged polymerase and a fluorescently modified nucleotide as the nucleotide becomes incorporated into the de novo chain. Other sequencing-by-synthesis technologies also exist.

[00157] Any suitable sequencing-by-synthesis platform can be used to identify mutations. As described above, four major sequencing-by-synthesis platforms are currently available: the Genome Sequencers from Roche/454 Life Sciences, the 1G Analyzer from Illumina/Solexa, the SOLiD system from Applied BioSystems, and the Heliscope system

from Helicos Biosciences. Sequencing-by-synthesis platforms have also been described by Pacific BioSciences and VisiGen Biotechnologies. In some embodiments, a plurality of nucleic acid molecules being sequenced is bound to a support (e.g., solid support). To immobilize the nucleic acid on a support, a capture sequence/universal priming site can be added at the 3' and/or 5' end of the template. The nucleic acids can be bound to the support by hybridizing the capture sequence to a complementary sequence covalently attached to the support. The capture sequence (also referred to as a universal capture sequence) is a nucleic acid sequence complementary to a sequence attached to a support that may dually serve as a universal primer.

[00158] As an alternative to a capture sequence, a member of a coupling pair (such as, e.g., antibody/antigen, receptor/ligand, or the avidin-biotin pair as described in, e.g., US Patent Application No. 2006/0252077) can be linked to each fragment to be captured on a surface coated with a respective second member of that coupling pair.

[00159] Subsequent to the capture, the sequence can be analyzed, for example, by single molecule detection/sequencing, e.g., as described in the Examples and in U.S. Pat. No. 7,283,337, including template-dependent sequencing-by-synthesis. In sequencing-by-synthesis, the surface-bound molecule is exposed to a plurality of labeled nucleotide triphosphates in the presence of polymerase. The sequence of the template is determined by the order of labeled nucleotides incorporated into the 3' end of the growing chain. This can be done in real time or can be done in a step-and-repeat mode. For real-time analysis, different optical labels to each nucleotide can be incorporated and multiple lasers can be utilized for stimulation of incorporated nucleotides.

[00160] Sequencing can also include other massively parallel sequencing or next generation sequencing (NGS) techniques and platforms. Additional examples of massively parallel sequencing techniques and platforms are the Illumina HiSeq or MiSeq, Thermo PGM or Proton, the Pac Bio RS II or Sequel, Qiagen's Gene Reader, and the Oxford Nanopore MinION. Additional similar current massively parallel sequencing technologies can be used, as well as future generations of these technologies.

[00161] Any cell type or tissue can be utilized to obtain nucleic acid samples for use in methods described herein. For example, a DNA or RNA sample can be obtained from a tumor or a bodily fluid, e.g., blood, obtained by known techniques (e.g. venipuncture) or saliva. Alternatively, nucleic acid tests can be performed on dry samples (e.g. hair or skin). In addition, a sample can be obtained for sequencing from a tumor and another sample

can be obtained from normal tissue for sequencing where the normal tissue is of the same tissue type as the tumor. A sample can be obtained for sequencing from a tumor and another sample can be obtained from normal tissue for sequencing where the normal tissue is of a distinct tissue type relative to the tumor.

[00162] Tumors can include one or more of lung cancer, melanoma, breast cancer, ovarian cancer, prostate cancer, kidney cancer, gastric cancer, colon cancer, testicular cancer, head and neck cancer, pancreatic cancer, brain cancer, B-cell lymphoma, acute myelogenous leukemia, chronic myelogenous leukemia, chronic lymphocytic leukemia, and T-cell lymphocytic leukemia, non-small cell lung cancer, and small cell lung cancer.

[00163] Alternatively, protein mass spectrometry can be used to identify or validate the presence of mutated peptides bound to MHC proteins on tumor cells. Peptides can be acid-eluted from tumor cells or from HLA molecules that are immunoprecipitated from tumor, and then identified using mass spectrometry.

IV. Neoantigens

[00164] Neoantigens can include nucleotides or polypeptides. For example, a neoantigen can be an RNA sequence that encodes for a polypeptide sequence. Neoantigens useful in vaccines can therefore include nucleotide sequences or polypeptide sequences.

[00165] Disclosed herein are isolated peptides that comprise tumor specific mutations identified by the methods disclosed herein, peptides that comprise known tumor specific mutations, and mutant polypeptides or fragments thereof identified by methods disclosed herein. Neoantigen peptides can be described in the context of their coding sequence where a neoantigen includes the nucleotide sequence (e.g., DNA or RNA) that codes for the related polypeptide sequence.

[00166] One or more polypeptides encoded by a neoantigen nucleotide sequence can comprise at least one of: a binding affinity with MHC with an IC50 value of less than 1000nM, for MHC Class I peptides a length of 8-15, 8, 9, 10, 11, 12, 13, 14, or 15 amino acids, presence of sequence motifs within or near the peptide promoting proteasome cleavage, and presence or sequence motifs promoting TAP transport. For MHC Class II peptides a length 6-30, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 amino acids, presence of sequence motifs within or near the peptide promoting cleavage by extracellular or lysosomal proteases (e.g., cathepsins) or HLA-DM catalyzed HLA binding.

[00167] One or more neoantigens can be presented on the surface of a tumor.

[00168] One or more neoantigens can be is immunogenic in a subject having a tumor, e.g., capable of eliciting a T-cell response or a B cell response in the subject.

[00169] One or more neoantigens that induce an autoimmune response in a subject can be excluded from consideration in the context of vaccine generation for a subject having a tumor.

[00170] The size of at least one neoantigenic peptide molecule can comprise, but is not limited to, about 5, about 6, about 7, about 8, about 9, about 10, about 11, about 12, about 13, about 14, about 15, about 16, about 17, about 18, about 19, about 20, about 21, about 22, about 23, about 24, about 25, about 26, about 27, about 28, about 29, about 30, about 31, about 32, about 33, about 34, about 35, about 36, about 37, about 38, about 39, about 40, about 41, about 42, about 43, about 44, about 45, about 46, about 47, about 48, about 49, about 50, about 60, about 70, about 80, about 90, about 100, about 110, about 120 or greater amino molecule residues, and any range derivable therein. In specific embodiments the neoantigenic peptide molecules are equal to or less than 50 amino acids.

[00171] Neoantigenic peptides and polypeptides can be: for MHC Class I 15 residues or less in length and usually consist of between about 8 and about 11 residues, particularly 9 or 10 residues; for MHC Class II, 6-30 residues, inclusive.

[00172] If desirable, a longer peptide can be designed in several ways. In one case, when presentation likelihoods of peptides on HLA alleles are predicted or known, a longer peptide could consist of either: (1) individual presented peptides with an extensions of 2-5 amino acids toward the N- and C-terminus of each corresponding gene product; (2) a concatenation of some or all of the presented peptides with extended sequences for each. In another case, when sequencing reveals a long (>10 residues) neoepitope sequence present in the tumor (e.g. due to a frameshift, read-through or intron inclusion that leads to a novel peptide sequence), a longer peptide would consist of: (3) the entire stretch of novel tumor-specific amino acids--thus bypassing the need for computational or in vitro test-based selection of the strongest HLA-presented shorter peptide. In both cases, use of a longer peptide allows endogenous processing by patient-cells and may lead to more effective antigen presentation and induction of T-cell responses.

[00173] Neoantigenic peptides and polypeptides can be presented on an HLA protein. In some aspects neoantigenic peptides and polypeptides are presented on an HLA protein with greater affinity than a wild-type peptide. In some aspects, a neoantigenic peptide or polypeptide can have an IC50 of at least less than 5000 nM, at least less than 1000 nM, at

least less than 500 nM, at least less than 250 nM, at least less than 200 nM, at least less than 150 nM, at least less than 100 nM, at least less than 50 nM or less.

[00174] In some aspects, neoantigenic peptides and polypeptides do not induce an autoimmune response and/or invoke immunological tolerance when administered to a subject.

[00175] Also provided are compositions comprising at least two or more neoantigenic peptides. In some embodiments the composition contains at least two distinct peptides. At least two distinct peptides can be derived from the same polypeptide. By distinct polypeptides is meant that the peptide vary by length, amino acid sequence, or both. The peptides are derived from any polypeptide known to or have been found to contain a tumor specific mutation. Suitable polypeptides from which the neoantigenic peptides can be derived can be found for example in the COSMIC database. COSMIC curates comprehensive information on somatic mutations in human cancer. The peptide contains the tumor specific mutation. In some aspects the tumor specific mutation is a driver mutation for a particular cancer type.

[00176] Neoantigenic peptides and polypeptides having a desired activity or property can be modified to provide certain desired attributes, e.g., improved pharmacological characteristics, while increasing or at least retaining substantially all of the biological activity of the unmodified peptide to bind the desired MHC molecule and activate the appropriate T-cell. For instance, neoantigenic peptide and polypeptides can be subject to various changes, such as substitutions, either conservative or non-conservative, where such changes might provide for certain advantages in their use, such as improved MHC binding, stability or presentation. By conservative substitutions is meant replacing an amino acid residue with another which is biologically and/or chemically similar, e.g., one hydrophobic residue for another, or one polar residue for another. The substitutions include combinations such as Gly, Ala; Val, Ile, Leu, Met; Asp, Glu; Asn, Gln; Ser, Thr; Lys, Arg; and Phe, Tyr. The effect of single amino acid substitutions may also be probed using D-amino acids. Such modifications can be made using well known peptide synthesis procedures, as described in e.g., Merrifield, Science 232:341-347 (1986), Barany & Merrifield, The Peptides, Gross & Meienhofer, eds. (N.Y., Academic Press), pp. 1-284 (1979); and Stewart & Young, Solid Phase Peptide Synthesis, (Rockford, Ill., Pierce), 2d Ed. (1984).

[00177] Modifications of peptides and polypeptides with various amino acid mimetics or unnatural amino acids can be particularly useful in increasing the stability of the peptide and

polypeptide in vivo. Stability can be assayed in a number of ways. For instance, peptidases and various biological media, such as human plasma and serum, have been used to test stability. See, e.g., Verhoef et al., Eur. J. Drug Metab Pharmacokin. 11:291-302 (1986). Half-life of the peptides can be conveniently determined using a 25% human serum (v/v) assay. The protocol is generally as follows. Pooled human serum (Type AB, non-heat inactivated) is delipidated by centrifugation before use. The serum is then diluted to 25% with RPMI tissue culture media and used to test peptide stability. At predetermined time intervals a small amount of reaction solution is removed and added to either 6% aqueous trichloracetic acid or ethanol. The cloudy reaction sample is cooled (4 degrees C) for 15 minutes and then spun to pellet the precipitated serum proteins. The presence of the peptides is then determined by reversed-phase HPLC using stability-specific chromatography conditions.

[00178] The peptides and polypeptides can be modified to provide desired attributes other than improved serum half-life. For instance, the ability of the peptides to induce CTL activity can be enhanced by linkage to a sequence which contains at least one epitope that is capable of inducing a T helper cell response. Immunogenic peptides/T helper conjugates can be linked by a spacer molecule. The spacer is typically comprised of relatively small, neutral molecules, such as amino acids or amino acid mimetics, which are substantially uncharged under physiological conditions. The spacers are typically selected from, e.g., Ala, Gly, or other neutral spacers of nonpolar amino acids or neutral polar amino acids. It will be understood that the optionally present spacer need not be comprised of the same residues and thus can be a hetero- or homo-oligomer. When present, the spacer will usually be at least one or two residues, more usually three to six residues. Alternatively, the peptide can be linked to the T helper peptide without a spacer.

[00179] A neoantigenic peptide can be linked to the T helper peptide either directly or via a spacer either at the amino or carboxy terminus of the peptide. The amino terminus of either the neoantigenic peptide or the T helper peptide can be acylated. Exemplary T helper peptides include tetanus toxoid 830-843, influenza 307-319, malaria circumsporozoite 382-398 and 378-389.

[00180] Proteins or peptides can be made by any technique known to those of skill in the art, including the expression of proteins, polypeptides or peptides through standard molecular biological techniques, the isolation of proteins or peptides from natural sources, or the chemical synthesis of proteins or peptides. The nucleotide and protein, polypeptide and peptide sequences corresponding to various genes have been previously disclosed, and can be

found at computerized databases known to those of ordinary skill in the art. One such database is the National Center for Biotechnology Information's Genbank and GenPept databases located at the National Institutes of Health website. The coding regions for known genes can be amplified and/or expressed using the techniques disclosed herein or as would be known to those of ordinary skill in the art. Alternatively, various commercial preparations of proteins, polypeptides and peptides are known to those of skill in the art.

[00181] In a further aspect a neoantigen includes a nucleic acid (e.g. polynucleotide) that encodes a neoantigenic peptide or portion thereof. The polynucleotide can be, e.g., DNA, cDNA, PNA, CNA, RNA (e.g., mRNA), either single- and/or double-stranded, or native or stabilized forms of polynucleotides, such as, e.g., polynucleotides with a phosphorothiate backbone, or combinations thereof and it may or may not contain introns. A still further aspect provides an expression vector capable of expressing a polypeptide or portion thereof. Expression vectors for different-cell types are well known in the art and can be selected without undue experimentation. Generally, DNA is inserted into an expression vector, such as a plasmid, in proper orientation and correct reading frame for expression. If necessary, DNA can be linked to the appropriate transcriptional and translational regulatory control nucleotide sequences recognized by the desired host, although such controls are generally available in the expression vector. The vector is then introduced into the host through standard techniques. Guidance can be found e.g. in Sambrook et al. (1989) Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

IV. Vaccine Compositions

[00182] Also disclosed herein is an immunogenic composition, e.g., a vaccine composition, capable of raising a specific immune response, e.g., a tumor-specific immune response. Vaccine compositions typically comprise a plurality of neoantigens, e.g., selected using a method described herein. Vaccine compositions can also be referred to as vaccines.

[00183] A vaccine can contain between 1 and 30 peptides, 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, or 30 different peptides, 6, 7, 8, 9, 10 11, 12, 13, or 14 different peptides, or 12, 13 or 14 different peptides. Peptides can include post-translational modifications. A vaccine can contain between 1 and 100 or more nucleotide sequences, 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, 100 or more different nucleotide sequences, 6, 7, 8, 9, 10 11, 12, 13, or 14 different nucleotide sequences, or 12, 13 or 14 different nucleotide sequences. A vaccine can contain between 1 and 30 neoantigen sequences, 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, 100 or more different neoantigen sequences, 6, 7, 8, 9, 10 11, 12, 13, or 14 different neoantigen sequences, or 12, 13 or 14 different neoantigen sequences.

[00184] In one embodiment, different peptides and/or polypeptides or nucleotide sequences encoding them are selected so that the peptides and/or polypeptides capable of associating with different MHC molecules, such as different MHC class I molecules and/or different MHC class II molecules. In some aspects, one vaccine composition comprises coding sequence for peptides and/or polypeptides capable of associating with the most frequently occurring MHC class I molecules and/or MHC class II molecules. Hence, vaccine compositions can comprise different fragments capable of associating with at least 2 preferred, at least 3 preferred, or at least 4 preferred MHC class I molecules and/or MHC class II molecules.

[00185] The vaccine composition can be capable of raising a specific cytotoxic T-cells response and/or a specific helper T-cell response.

[00186] A vaccine composition can further comprise an adjuvant and/or a carrier. Examples of useful adjuvants and carriers are given herein below. A composition can be associated with a carrier such as e.g. a protein or an antigen-presenting cell such as e.g. a dendritic cell (DC) capable of presenting the peptide to a T-cell.

[00187] Adjuvants are any substance whose admixture into a vaccine composition increases or otherwise modifies the immune response to a neoantigen. Carriers can be scaffold structures, for example a polypeptide or a polysaccharide, to which a neoantigen, is capable of being associated. Optionally, adjuvants are conjugated covalently or non-covalently.

[00188] The ability of an adjuvant to increase an immune response to an antigen is typically manifested by a significant or substantial increase in an immune-mediated reaction, or reduction in disease symptoms. For example, an increase in humoral immunity is typically manifested by a significant increase in the titer of antibodies raised to the antigen, and an increase in T-cell activity is typically manifested in increased cell proliferation, or cellular

cytotoxicity, or cytokine secretion. An adjuvant may also alter an immune response, for example, by changing a primarily humoral or Th response into a primarily cellular, or Th response.

[00189] Suitable adjuvants include, but are not limited to 1018 ISS, alum, aluminium salts, Amplivax, AS15, BCG, CP-870,893, CpG7909, CyaA, dSLIM, GM-CSF, IC30, IC31, Imiquimod, ImuFact IMP321, IS Patch, ISS, ISCOMATRIX, JuvImmune, LipoVac, MF59, monophosphoryl lipid A, Montanide IMS 1312, Montanide ISA 206, Montanide ISA 50V, Montanide ISA-51, OK-432, OM-174, OM-197-MP-EC, ONTAK, PepTel vector system, PLG microparticles, resiquimod, SRL172, Virosomes and other Virus-like particles, YF-17D, VEGF trap, R848, beta-glucan, Pam3Cys, Aquila's QS21 stimulon (Aquila Biotech, Worcester, Mass., USA) which is derived from saponin, mycobacterial extracts and synthetic bacterial cell wall mimics, and other proprietary adjuvants such as Ribi's Detox. Quil or Superfos. Adjuvants such as incomplete Freund's or GM-CSF are useful. Several immunological adjuvants (e.g., MF59) specific for dendritic cells and their preparation have been described previously (Dupuis M, et al., Cell Immunol. 1998; 186(1):18-27; Allison A C; Dev Biol Stand. 1998; 92:3-11). Also cytokines can be used. Several cytokines have been directly linked to influencing dendritic cell migration to lymphoid tissues (e.g., TNF-alpha), accelerating the maturation of dendritic cells into efficient antigen-presenting cells for T-lymphocytes (e.g., GM-CSF, IL-1 and IL-4) (U.S. Pat. No. 5,849,589, specifically incorporated herein by reference in its entirety) and acting as immunoadjuvants (e.g., IL-12) (Gabrilovich D I, et al., J Immunother Emphasis Tumor Immunol. 1996 (6):414-418).

[00190] CpG immunostimulatory oligonucleotides have also been reported to enhance the effects of adjuvants in a vaccine setting. Other TLR binding molecules such as RNA binding TLR 7, TLR 8 and/or TLR 9 may also be used.

[00191] Other examples of useful adjuvants include, but are not limited to, chemically modified CpGs (e.g. CpR, Idera), Poly(I:C)(e.g. polyi:CI2U), non-CpG bacterial DNA or RNA as well as immunoactive small molecules and antibodies such as cyclophosphamide, sunitinib, bevacizumab, celebrex, NCX-4016, sildenafil, tadalafil, vardenafil, sorafinib, XL-999, CP-547632, pazopanib, ZD2171, AZD2171, ipilimumab, tremelimumab, and SC58175, which may act therapeutically and/or as an adjuvant. The amounts and concentrations of adjuvants and additives can readily be determined by the skilled artisan without undue experimentation. Additional adjuvants include colony-stimulating factors, such as Granulocyte Macrophage Colony Stimulating Factor (GM-CSF, sargramostim).

[00192] A vaccine composition can comprise more than one different adjuvant. Furthermore, a therapeutic composition can comprise any adjuvant substance including any of the above or combinations thereof. It is also contemplated that a vaccine and an adjuvant can be administered together or separately in any appropriate sequence.

[00193] A carrier (or excipient) can be present independently of an adjuvant. The function of a carrier can for example be to increase the molecular weight of in particular mutant to increase activity or immunogenicity, to confer stability, to increase the biological activity, or to increase serum half-life. Furthermore, a carrier can aid presenting peptides to T-cells. A carrier can be any suitable carrier known to the person skilled in the art, for example a protein or an antigen presenting cell. A carrier protein could be but is not limited to keyhole limpet hemocyanin, serum proteins such as transferrin, bovine serum albumin, human serum albumin, thyroglobulin or ovalbumin, immunoglobulins, or hormones, such as insulin or palmitic acid. For immunization of humans, the carrier is generally a physiologically acceptable carrier acceptable to humans and safe. However, tetanus toxoid and/or diptheria toxoid are suitable carriers. Alternatively, the carrier can be dextrans for example sepharose.

[00194] Cytotoxic T-cells (CTLs) recognize an antigen in the form of a peptide bound to an MHC molecule rather than the intact foreign antigen itself. The MHC molecule itself is located at the cell surface of an antigen presenting cell. Thus, an activation of CTLs is possible if a trimeric complex of peptide antigen, MHC molecule, and APC is present. Correspondingly, it may enhance the immune response if not only the peptide is used for activation of CTLs, but if additionally APCs with the respective MHC molecule are added. Therefore, in some embodiments a vaccine composition additionally contains at least one antigen presenting cell.

[00195] Neoantigens can also be included in viral vector-based vaccine platforms, such as vaccinia, fowlpox, self-replicating alphavirus, marabavirus, adenovirus (*See, e.g.,* Tatsis et al., Adenoviruses, *Molecular Therapy* (2004) 10, 616—629), or lentivirus, including but not limited to second, third or hybrid second/third generation lentivirus and recombinant lentivirus of any generation designed to target specific cell types or receptors (*See, e.g.,* Hu et al., Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases, *Immunol Rev.* (2011) 239(1): 45-61, Sakuma et al., Lentiviral vectors: basic to translational, *Biochem J.* (2012) 443(3):603-18, Cooper et al., Rescue of splicing-mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter, *Nucl. Acids Res.* (2015) 43 (1): 682-690, Zufferey et al., Self-Inactivating Lentivirus Vector for

Safe and Efficient In Vivo Gene Delivery, *J. Virol.* (1998) 72 (12): 9873-9880). Dependent on the packaging capacity of the above mentioned viral vector-based vaccine platforms, this approach can deliver one or more nucleotide sequences that encode one or more neoantigen peptides. The sequences may be flanked by non-mutated sequences, may be separated by linkers or may be preceded with one or more sequences targeting a subcellular compartment (See, e.g., Gros et al., Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients, *Nat Med.* (2016) 22 (4):433-8, Stronen et al., Targeting of cancer neoantigens with donor-derived T-cell receptor repertoires, *Science.* (2016) 352 (6291):1337-41, Lu et al., Efficient identification of mutated cancer antigens recognized by T-cells associated with durable tumor regressions, *Clin Cancer Res.* (2014) 20(13):3401-10). Upon introduction into a host, infected cells express the neoantigens, and thereby elicit a host immune (e.g., CTL) response against the peptide(s). Vaccinia vectors and methods useful in immunization protocols are described in, e.g., U.S. Pat. No. 4,722,848. Another vector is BCG (Bacille Calmette Guerin). BCG vectors are described in Stover et al. (Nature 351:456-460 (1991)). A wide variety of other vaccine vectors useful for therapeutic administration or immunization of neoantigens, e.g., Salmonella typhi vectors, and the like will be apparent to those skilled in the art from the description herein.

IV.A. Additional Considerations for Vaccine Design and Manufacture

IV.A.1. Determination of a set of peptides that cover all tumor subclones

[00196] Truncal peptides, meaning those presented by all or most tumor subclones, will be prioritized for inclusion into the vaccine.[53] Optionally, if there are no truncal peptides predicted to be presented and immunogenic with high probability, or if the number of truncal peptides predicted to be presented and immunogenic with high probability is small enough that additional non-truncal peptides can be included in the vaccine, then further peptides can be prioritized by estimating the number and identity of tumor subclones and choosing peptides so as to maximize the number of tumor subclones covered by the vaccine.[54]

IV.A.2. Neoantigen prioritization

[00197] After all of the above neoantigen filters are applied, more candidate neoantigens may still be available for vaccine inclusion than the vaccine technology can support. Additionally, uncertainty about various aspects of the neoantigen analysis may remain and tradeoffs may exist between different properties of candidate vaccine neoantigens. Thus, in

place of predetermined filters at each step of the selection process, an integrated multi-dimensional model can be considered that places candidate neoantigens in a space with at least the following axes and optimizes selection using an integrative approach.

1. Risk of auto-immunity or tolerance (risk of germline) (lower risk of auto-immunity is typically preferred)
2. Probability of sequencing artifact (lower probability of artifact is typically preferred)
3. Probability of immunogenicity (higher probability of immunogenicity is typically preferred)
4. Probability of presentation (higher probability of presentation is typically preferred)
5. Gene expression (higher expression is typically preferred)
6. Coverage of HLA genes (larger number of HLA molecules involved in the presentation of a set of neoantigens may lower the probability that a tumor will escape immune attack via downregulation or mutation of HLA molecules)
7. Coverage of HLA classes (covering both HLA-I and HLA-II may increase the probability of therapeutic response and decrease the probability of tumor escape)

V. Therapeutic and Manufacturing Methods

[00198] Also provided is a method of inducing a tumor specific immune response in a subject, vaccinating against a tumor, treating and or alleviating a symptom of cancer in a subject by administering to the subject one or more neoantigens such as a plurality of neoantigens identified using methods disclosed herein.

[00199] In some aspects, a subject has been diagnosed with cancer or is at risk of developing cancer. A subject can be a human, dog, cat, horse or any animal in which a tumor specific immune response is desired. A tumor can be any solid tumor such as breast, ovarian, prostate, lung, kidney, gastric, colon, testicular, head and neck, pancreas, brain, melanoma, and other tumors of tissue organs and hematological tumors, such as lymphomas and leukemias, including acute myelogenous leukemia, chronic myelogenous leukemia, chronic lymphocytic leukemia, T-cell lymphocytic leukemia, and B cell lymphomas.

[00200] A neoantigen can be administered in an amount sufficient to induce a CTL response.

[00201] A neoantigen can be administered alone or in combination with other therapeutic agents. The therapeutic agent is for example, a chemotherapeutic agent, radiation, or immunotherapy. Any suitable therapeutic treatment for a particular cancer can be administered.

[00202] In addition, a subject can be further administered an anti-immunosuppressive/immunostimulatory agent such as a checkpoint inhibitor. For example, the subject can be further administered an anti-CTLA antibody or anti-PD-1 or anti-PD-L1.

Blockade of CTLA-4 or PD-L1 by antibodies can enhance the immune response to cancerous cells in the patient. In particular, CTLA-4 blockade has been shown effective when following a vaccination protocol.

[00203] The optimum amount of each neoantigen to be included in a vaccine composition and the optimum dosing regimen can be determined. For example, a neoantigen or its variant can be prepared for intravenous (i.v.) injection, sub-cutaneous (s.c.) injection, intradermal (i.d.) injection, intraperitoneal (i.p.) injection, intramuscular (i.m.) injection. Methods of injection include s.c., i.d., i.p., i.m., and i.v. Methods of DNA or RNA injection include i.d., i.m., s.c., i.p. and i.v. Other methods of administration of the vaccine composition are known to those skilled in the art.

[00204] A vaccine can be compiled so that the selection, number and/or amount of neoantigens present in the composition is/are tissue, cancer, and/or patient-specific. For instance, the exact selection of peptides can be guided by expression patterns of the parent proteins in a given tissue. The selection can be dependent on the specific type of cancer, the status of the disease, earlier treatment regimens, the immune status of the patient, and, of course, the HLA-haplotype of the patient. Furthermore, a vaccine can contain individualized components, according to personal needs of the particular patient. Examples include varying the selection of neoantigens according to the expression of the neoantigen in the particular patient or adjustments for secondary treatments following a first round or scheme of treatment.

[00205] For a composition to be used as a vaccine for cancer, neoantigens with similar normal self-peptides that are expressed in high amounts in normal tissues can be avoided or be present in low amounts in a composition described herein. On the other hand, if it is known that the tumor of a patient expresses high amounts of a certain neoantigen, the respective pharmaceutical composition for treatment of this cancer can be present in high amounts and/or more than one neoantigen specific for this particularly neoantigen or pathway of this neoantigen can be included.

[00206] Compositions comprising a neoantigen can be administered to an individual already suffering from cancer. In therapeutic applications, compositions are administered to a patient in an amount sufficient to elicit an effective CTL response to the tumor antigen and to cure or at least partially arrest symptoms and/or complications. An amount adequate to accomplish this is defined as "therapeutically effective dose." Amounts effective for this use will depend on, e.g., the composition, the manner of administration, the stage and severity of

the disease being treated, the weight and general state of health of the patient, and the judgment of the prescribing physician. It should be kept in mind that compositions can generally be employed in serious disease states, that is, life-threatening or potentially life threatening situations, especially when the cancer has metastasized. In such cases, in view of the minimization of extraneous substances and the relative nontoxic nature of a neoantigen, it is possible and can be felt desirable by the treating physician to administer substantial excesses of these compositions.

[00207] For therapeutic use, administration can begin at the detection or surgical removal of tumors. This is followed by boosting doses until at least symptoms are substantially abated and for a period thereafter.

[00208] The pharmaceutical compositions (e.g., vaccine compositions) for therapeutic treatment are intended for parenteral, topical, nasal, oral or local administration. A pharmaceutical compositions can be administered parenterally, e.g., intravenously, subcutaneously, intradermally, or intramuscularly. The compositions can be administered at the site of surgical excision to induce a local immune response to the tumor. Disclosed herein are compositions for parenteral administration which comprise a solution of the neoantigen and vaccine compositions are dissolved or suspended in an acceptable carrier, e.g., an aqueous carrier. A variety of aqueous carriers can be used, e.g., water, buffered water, 0.9% saline, 0.3% glycine, hyaluronic acid and the like. These compositions can be sterilized by conventional, well known sterilization techniques, or can be sterile filtered. The resulting aqueous solutions can be packaged for use as is, or lyophilized, the lyophilized preparation being combined with a sterile solution prior to administration. The compositions may contain pharmaceutically acceptable auxiliary substances as required to approximate physiological conditions, such as pH adjusting and buffering agents, tonicity adjusting agents, wetting agents and the like, for example, sodium acetate, sodium lactate, sodium chloride, potassium chloride, calcium chloride, sorbitan monolaurate, triethanolamine oleate, etc.

[00209] Neoantigens can also be administered via liposomes, which target them to a particular cells tissue, such as lymphoid tissue. Liposomes are also useful in increasing half-life. Liposomes include emulsions, foams, micelles, insoluble monolayers, liquid crystals, phospholipid dispersions, lamellar layers and the like. In these preparations the neoantigen to be delivered is incorporated as part of a liposome, alone or in conjunction with a molecule which binds to, e.g., a receptor prevalent among lymphoid cells, such as monoclonal antibodies which bind to the CD45 antigen, or with other therapeutic or immunogenic

compositions. Thus, liposomes filled with a desired neoantigen can be directed to the site of lymphoid cells, where the liposomes then deliver the selected therapeutic/immunogenic compositions. Liposomes can be formed from standard vesicle-forming lipids, which generally include neutral and negatively charged phospholipids and a sterol, such as cholesterol. The selection of lipids is generally guided by consideration of, e.g., liposome size, acid lability and stability of the liposomes in the blood stream. A variety of methods are available for preparing liposomes, as described in, e.g., Szoka et al., Ann. Rev. Biophys. Bioeng. 9; 467 (1980), U.S. Pat. Nos. 4,235,871, 4,501,728, 4,501,728, 4,837,028, and 5,019,369.

[00210] For targeting to the immune cells, a ligand to be incorporated into the liposome can include, e.g., antibodies or fragments thereof specific for cell surface determinants of the desired immune system cells. A liposome suspension can be administered intravenously, locally, topically, etc. in a dose which varies according to, inter alia, the manner of administration, the peptide being delivered, and the stage of the disease being treated.

[00211] For therapeutic or immunization purposes, nucleic acids encoding a peptide and optionally one or more of the peptides described herein can also be administered to the patient. A number of methods are conveniently used to deliver the nucleic acids to the patient. For instance, the nucleic acid can be delivered directly, as "naked DNA". This approach is described, for instance, in Wolff et al., Science 247: 1465-1468 (1990) as well as U.S. Pat. Nos. 5,580,859 and 5,589,466. The nucleic acids can also be administered using ballistic delivery as described, for instance, in U.S. Pat. No. 5,204,253. Particles comprised solely of DNA can be administered. Alternatively, DNA can be adhered to particles, such as gold particles. Approaches for delivering nucleic acid sequences can include viral vectors, mRNA vectors, and DNA vectors with or without electroporation.

[00212] The nucleic acids can also be delivered complexed to cationic compounds, such as cationic lipids. Lipid-mediated gene delivery methods are described, for instance, in 9618372WOAWO 96/18372; 9324640WOAWO 93/24640; Mannino & Gould-Fogerite, BioTechniques 6(7): 682-691 (1988); U.S. Pat. No. 5,279,833 Rose U.S. Pat. No. 5,279,833; 9106309WOAWO 91/06309; and Felgner et al., Proc. Natl. Acad. Sci. USA 84: 7413-7414 (1987).

[00213] Neoantigens can also be included in viral vector-based vaccine platforms, such as vaccinia, fowlpox, self-replicating alphavirus, marabavirus, adenovirus (*See, e.g.,* Tatsis et al., Adenoviruses, *Molecular Therapy* (2004) 10, 616—629), or lentivirus, including but not

limited to second, third or hybrid second/third generation lentivirus and recombinant lentivirus of any generation designed to target specific cell types or receptors (*See, e.g.,* Hu et al., Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases, *Immunol Rev.* (2011) 239(1): 45-61, Sakuma et al., Lentiviral vectors: basic to translational, *Biochem J.* (2012) 443(3):603-18, Cooper et al., Rescue of splicing-mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter, *Nucl. Acids Res.* (2015) 43 (1): 682-690, Zufferey et al., Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery, *J. Virol.* (1998) 72 (12): 9873-9880). Dependent on the packaging capacity of the above mentioned viral vector-based vaccine platforms, this approach can deliver one or more nucleotide sequences that encode one or more neoantigen peptides. The sequences may be flanked by non-mutated sequences, may be separated by linkers or may be preceded with one or more sequences targeting a subcellular compartment (See, e.g., Gros et al., Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients, *Nat Med.* (2016) 22 (4):433-8, Stronen et al., Targeting of cancer neoantigens with donor-derived T-cell receptor repertoires, *Science.* (2016) 352 (6291):1337-41, Lu et al., Efficient identification of mutated cancer antigens recognized by T-cells associated with durable tumor regressions, *Clin Cancer Res.* (2014) 20(13):3401-10). Upon introduction into a host, infected cells express the neoantigens, and thereby elicit a host immune (e.g., CTL) response against the peptide(s). Vaccinia vectors and methods useful in immunization protocols are described in, e.g., U.S. Pat. No. 4,722,848. Another vector is BCG (Bacille Calmette Guerin). BCG vectors are described in Stover et al. (Nature 351:456-460 (1991)). A wide variety of other vaccine vectors useful for therapeutic administration or immunization of neoantigens, e.g., Salmonella typhi vectors, and the like will be apparent to those skilled in the art from the description herein.

[00214] A means of administering nucleic acids uses minigene constructs encoding one or multiple epitopes. To create a DNA sequence encoding the selected CTL epitopes (minigene) for expression in human cells, the amino acid sequences of the epitopes are reverse translated. A human codon usage table is used to guide the codon choice for each amino acid. These epitope-encoding DNA sequences are directly adjoined, creating a continuous polypeptide sequence. To optimize expression and/or immunogenicity, additional elements can be incorporated into the minigene design. Examples of amino acid sequence that could be reverse translated and included in the minigene sequence include: helper T lymphocyte, epitopes, a leader (signal) sequence, and an endoplasmic reticulum retention signal. In

addition, MHC presentation of CTL epitopes can be improved by including synthetic (e.g. poly-alanine) or naturally-occurring flanking sequences adjacent to the CTL epitopes. The minigene sequence is converted to DNA by assembling oligonucleotides that encode the plus and minus strands of the minigene. Overlapping oligonucleotides (30-100 bases long) are synthesized, phosphorylated, purified and annealed under appropriate conditions using well known techniques. The ends of the oligonucleotides are joined using T4 DNA ligase. This synthetic minigene, encoding the CTL epitope polypeptide, can then cloned into a desired expression vector.

[00215] Purified plasmid DNA can be prepared for injection using a variety of formulations. The simplest of these is reconstitution of lyophilized DNA in sterile phosphate-buffer saline (PBS). A variety of methods have been described, and new techniques can become available. As noted above, nucleic acids are conveniently formulated with cationic lipids. In addition, glycolipids, fusogenic liposomes, peptides and compounds referred to collectively as protective, interactive, non-condensing (PINC) could also be complexed to purified plasmid DNA to influence variables such as stability, intramuscular dispersion, or trafficking to specific organs or cell types.

[00216] Also disclosed is a method of manufacturing a tumor vaccine, comprising performing the steps of a method disclosed herein; and producing a tumor vaccine comprising a plurality of neoantigens or a subset of the plurality of neoantigens.

[00217] Neoantigens disclosed herein can be manufactured using methods known in the art. For example, a method of producing a neoantigen or a vector (e.g., a vector including at least one sequence encoding one or more neoantigens) disclosed herein can include culturing a host cell under conditions suitable for expressing the neoantigen or vector wherein the host cell comprises at least one polynucleotide encoding the neoantigen or vector, and purifying the neoantigen or vector. Standard purification methods include chromatographic techniques, electrophoretic, immunological, precipitation, dialysis, filtration, concentration, and chromatofocusing techniques.

[00218] Host cells can include a Chinese Hamster Ovary (CHO) cell, NS0 cell, yeast, or a HEK293 cell. Host cells can be transformed with one or more polynucleotides comprising at least one nucleic acid sequence that encodes a neoantigen or vector disclosed herein, optionally wherein the isolated polynucleotide further comprises a promoter sequence operably linked to the at least one nucleic acid sequence that encodes the neoantigen or vector. In certain embodiments the isolated polynucleotide can be cDNA.

VI. Neoantigen Identification

VI.A. Neoantigen Candidate Identification.

[00219] Research methods for NGS analysis of tumor and normal exome and transcriptomes have been described and applied in the neoantigen identification space. [6,14,15] The example below considers certain optimizations for greater sensitivity and specificity for neoantigen identification in the clinical setting. These optimizations can be grouped into two areas, those related to laboratory processes and those related to the NGS data analysis.

VI.A.1. Laboratory process optimizations

[00220] The process improvements presented here address challenges in high-accuracy neoantigen discovery from clinical specimens with low tumor content and small volumes by extending concepts developed for reliable cancer driver gene assessment in targeted cancer panels[16] to the whole- exome and -transcriptome setting necessary for neoantigen identification. Specifically, these improvements include:

1. Targeting deep (>500x) unique average coverage across the tumor exome to detect mutations present at low mutant allele frequency due to either low tumor content or subclonal state.
2. Targeting uniform coverage across the tumor exome, with <5% of bases covered at <100x, so that the fewest possible neoantigens are missed, by, for instance:
    a. Employing DNA-based capture probes with individual probe QC[17]
    b. Including additional baits for poorly covered regions
3. Targeting uniform coverage across the normal exome, where <5% of bases are covered at <20x so that the fewest neoantigens possible remain unclassified for somatic/germline status (and thus not usable as TSNAs)
4. To minimize the total amount of sequencing required, sequence capture probes will be designed for coding regions of genes only, as non-coding RNA cannot give rise to neoantigens. Additional optimizations include:
    a. supplementary probes for HLA genes, which are GC-rich and poorly captured by standard exome sequencing[18]
    b. exclusion of genes predicted to generate few or no candidate neoantigens, due to factors such as insufficient expression, suboptimal digestion by the proteasome, or unusual sequence features.

5. Tumor RNA will likewise be sequenced at high depth (>100M reads) in order to enable variant detection, quantification of gene and splice-variant ("isoform") expression, and fusion detection. RNA from FFPE samples will be extracted using probe-based enrichment[19], with the same or similar probes used to capture exomes in DNA.

VI.A.2. NGS data analysis optimizations

[00221] Improvements in analysis methods address the suboptimal sensitivity and specificity of common research mutation calling approaches, and specifically consider customizations relevant for neoantigen identification in the clinical setting. These include:

1. Using the HG38 reference human genome or a later version for alignment, as it contains multiple MHC regions assemblies better reflective of population polymorphism, in contrast to previous genome releases.
2. Overcoming the limitations of single variant callers [20] by merging results from different programs[5]

    a. Single-nucleotide variants and indels will be detected from tumor DNA, tumor RNA and normal DNA with a suite of tools including: programs based on comparisons of tumor and normal DNA, such as Strelka [21] and Mutect [22]; and programs that incorporate tumor DNA, tumor RNA and normal DNA, such as UNCeqR, which is particularly advantageous in low-purity samples [23].
    b. Indels will be determined with programs that perform local re-assembly, such as Strelka and ABRA [24].
    c. Structural rearrangements will be determined using dedicated tools such as Pindel [25] or Breakseq [26].
3. In order to detect and prevent sample swaps, variant calls from samples for the same patient will be compared at a chosen number of polymorphic sites.
4. Extensive filtering of artefactual calls will be performed, for instance, by:
    a. Removal of variants found in normal DNA, potentially with relaxed detection parameters in cases of low coverage, and with a permissive proximity criterion in case of indels
    b. Removal of variants due to low mapping quality or low base quality[27].

c. Removal of variants stemming from recurrent sequencing artifacts, even if not observed in the corresponding normal[27]. Examples include variants primarily detected on one strand.

d. Removal of variants detected in an unrelated set of controls[27]

5. Accurate HLA calling from normal exome using one of seq2HLA [28], ATHLATES [29] or Optitype and also combining exome and RNA sequencing data [28]. Additional potential optimizations include the adoption of a dedicated assay for HLA typing such as long-read DNA sequencing[30], or the adaptation of a method for joining RNA fragments to retain continuity [31].

6. Robust detection of neo-ORFs arising from tumor-specific splice variants will be performed by assembling transcripts from RNA-seq data using CLASS [32], Bayesembler [33], StringTie [34] or a similar program in its reference-guided mode (i.e., using known transcript structures rather than attempting to recreate transcripts in their entirety from each experiment). While Cufflinks [35] is commonly used for this purpose, it frequently produces implausibly large numbers of splice variants, many of them far shorter than the full-length gene, and can fail to recover simple positive controls. Coding sequences and nonsense-mediated decay potential will be determined with tools such as SpliceR[36] and MAMBA[37], with mutant sequences re-introduced. Gene expression will be determined with a tool such as Cufflinks[35] or Express (Roberts and Pachter, 2013). Wild-type and mutant-specific expression counts and/or relative levels will be determined with tools developed for these purposes, such as ASE[38] or HTSeq[39]. Potential filtering steps include:

a. Removal of candidate neo-ORFs deemed to be insufficiently expressed.

b. Removal of candidate neo-ORFs predicted to trigger non-sense mediated decay (NMD).

7. Candidate neoantigens observed only in RNA (e.g., neoORFs) that cannot directly be verified as tumor-specific will be categorized as likely tumor-specific according to additional parameters, for instance by considering:

a. Presence of supporting tumor DNA-only *cis*-acting frameshift or splice-site mutations

b. Presence of corroborating tumor DNA-only trans-acting mutation in a splicing factor. For instance, in three independently published experiments with R625-mutant SF3B1, the genes exhibiting the most differentially splicing were

concordant even though one experiment examined uveal melanoma patients [40], the second a uveal melanoma cell line [41], and the third breast cancer patients [42].

c. For novel splicing isoforms, presence of corroborating "novel" splice-junction reads in the RNASeq data.
d. For novel re-arrangements, presence of corroborating juxta-exon reads in tumor DNA that are absent from normal DNA
e. Absence from gene expression compendium such as GTEx[43] (i.e. making germline origin less likely)

8. Complementing the reference genome alignment-based analysis by comparing assembled DNA tumor and normal reads (or k-mers from such reads) directly to avoid alignment and annotation based errors and artifacts. (e.g. for somatic variants arising near germline variants or repeat-context indels)

[00222] In samples with poly-adenylated RNA, the presence of viral and microbial RNA in the RNA-seq data will be assessed using RNA CoMPASS[44] or a similar method, toward the identification of additional factors that may predict patient response.

VI.B. Isolation and Detection of HLA Peptides

[00223] Isolation of HLA-peptide molecules was performed using classic immunoprecipitation (IP) methods after lysis and solubilization of the tissue sample[55-58]. A clarified lysate was used for HLA specific IP.

[00224] Immunoprecipitation was performed using antibodies coupled to beads where the antibody is specific for HLA molecules. For a pan-Class I HLA immunoprecipitation, a pan-Class I CR antibody is used, for Class II HLA – DR, an HLA-DR antibody is used. Antibody is covalently attached to NHS-sepharose beads during overnight incubation. After covalent attachment, the beads were washed and aliquoted for IP.[59, 60] Immunoprecipitations can also be performed with antibodies that are not covalently attached to beads. Typically this is done using sepharose or magnetic beads coated with Protein A and/or Protein G to hold the antibody to the column. Some antibodies that can be used to selectively enrich MHC/peptide complex are listed below.

Antibody Name	Specificity
W6/32	Class I HLA-A, B, C
L243	Class II – HLA-DR

Tu36	Class II – HLA-DR
LN3	Class II – HLA-DR
Tu39	Class II – HLA-DR, DP, DQ

[00225] The clarified tissue lysate is added to the antibody beads for the immunoprecipitation. After immunoprecipitation, the beads are removed from the lysate and the lysate stored for additional experiments, including additional IPs. The IP beads are washed to remove non-specific binding and the HLA/peptide complex is eluted from the beads using standard techniques. The protein components are removed from the peptides using a molecular weight spin column or C18 fractionation. The resultant peptides are taken to dryness by SpeedVac evaporation and in some instances are stored at -20C prior to MS analysis.

[00226] Dried peptides are reconstituted in an HPLC buffer suitable for reverse phase chromatography and loaded onto a C-18 microcapillary HPLC column for gradient elution in a Fusion Lumos mass spectrometer (Thermo). MS1 spectra of peptide mass/charge (m/z) were collected in the Orbitrap detector at high resolution followed by MS2 low resolution scans collected in the ion trap detector after HCD fragmentation of the selected ion. Additionally, MS2 spectra can be obtained using either CID or ETD fragmentation methods or any combination of the three techniques to attain greater amino acid coverage of the peptide. MS2 spectra can also be measured with high resolution mass accuracy in the Orbitrap detector.

[00227] MS2 spectra from each analysis are searched against a protein database using Comet[61, 62] and the peptide identification are scored using Percolator[63-65]. Additional sequencing is performed using PEAKS studio (Bioinformatics Solutions Inc.) and other search engines or sequencing methods can be used including spectral matching and *de novo* sequencing[75].

VI.B.1. MS limit of detection studies in support of comprehensive HLA peptide sequencing.

[00228] Using the peptide YVYVADVAAK it was determined what the limits of detection are using different amounts of peptide loaded onto the LC column. The amounts of peptide tested were 1 pmol, 100fmol, 10 fmol, 1 fmol, and 100amol. (Table 1) The results are shown in FIG. 1F. These results indicate that the lowest limit of detection (LoD) is in the attomol range (10^{-18}), that the dynamic range spans five orders of magnitude, and that the signal to noise appears sufficient for sequencing at low femtomol ranges (10^{-15}).

Peptide m/z	Loaded on Column	Copies/Cell in 1e9cells
566.830	1 pmol	600
562.823	100 fmol	60
559.816	10 fmol	6
556.810	1 fmol	0.6
553.802	100 amol	0.06

VII. Presentation Model

VII.A. System Overview

[00229] FIG. 2A is an overview of an environment 100 for identifying likelihoods of peptide presentation in patients, in accordance with an embodiment. The environment 100 provides context in order to introduce a presentation identification system 160, itself including a presentation information store 165.

[00230] The presentation identification system 160 is one or computer models, embodied in a computing system as discussed below with respect to FIG. 38, that receives peptide sequences associated with a set of MHC alleles and determines likelihoods that the peptide sequences will be presented by one or more of the set of associated MHC alleles. The presentation identification system 160 may be applied to both class I and class II MHC alleles. This is useful in a variety of contexts. One specific use case for the presentation identification system 160 is that it is able to receive nucleotide sequences of candidate neoantigens associated with a set of MHC alleles from tumor cells of a patient 110 and determine likelihoods that the candidate neoantigens will be presented by one or more of the associated MHC alleles of the tumor and/or induce immunogenic responses in the immune system of the patient 110. Those candidate neoantigens with high likelihoods as determined by system 160 can be selected for inclusion in a vaccine 118, such an anti-tumor immune response can be elicited from the immune system of the patient 110 providing the tumor cells. Additionally, T-cells with TCRs that are responsive to candidate neoantigens with high presentation likelihoods can be produced for use in T-cell therapy, thereby also eliciting an anti-tumor immune response from the immune system of the patient 110.

[00231] The presentation identification system 160 determines presentation likelihoods through one or more presentation models. Specifically, the presentation models generate likelihoods of whether given peptide sequences will be presented for a set of associated MHC alleles, and are generated based on presentation information stored in store 165. For example, the presentation models may generate likelihoods of whether a peptide sequence "YVYVADVAAK" will be presented for the set of alleles HLA-A*02:01, HLA-A*03:01, HLA-B*07:02, HLA-B*08:03, HLA-C*01:04 on the cell surface of the sample. As another example, the presentation models may also generate likelihoods of whether the peptide sequence "YVYVADVAAK" will be presented by HLA alleles having HLA allele sequences "AYANGPW", "UIIKNFDL", "WRTSAOGH". The presentation information 165 contains information on whether peptides bind to different types of MHC alleles such that those peptides are presented by MHC alleles, which in the models is determined depending on positions of amino acids in the peptide sequences. The presentation model can predict whether an unrecognized peptide sequence will be presented in association with an associated set of MHC alleles based on the presentation information 165. As previously mentioned, the presentation models may be applied to both class I and class II MHC alleles.

[00232] The term "HLA coverage" is used through this specification. As used throughout the specification, "HLA coverage" can be applied to an individual and/or to a population of individuals. As applied to an individual, "HLA coverage" refers to the proportion of HLA alleles found within the individual's genome for which a presentation model exists. For example, for a homozygous individual with HLA type A*02:01, A*02:01, B*07:02, B*07:02, C*07:02, C*07:02, if a presentation model exists for alleles A*02:01 and B*07:02, but not C*07:02, then the HLA coverage for the individual is 4/6.

[00233] As applied to a population of individuals, "HLA coverage" refers to the proportion of individuals in the population for each possible level of individual HLA coverage for which a presentation model exists. In the case of human individuals, each human genome contains six HLA alleles. Therefore possible levels of individual HLA coverages include 0/6, 1/6, 2/6,..., 6/6. Thus for example, in a population of individuals, if half of the individuals in the population have an individual HLA coverage of 2/6 and half of the individuals in the population have an individual HLA coverage of 6/6, then the HLA coverage of the population is 0% for individual HLA coverage 0/6, 0% for individual HLA coverage 1/6, 50% for individual HLA coverage 2/6, 0% for individual HLA coverage 3/6,

0% for individual HLA coverage 4/6, 0% for individual HLA coverage 5/6, and 50% for individual HLA coverage 6/6.

[00234] As described in further detail below with regard to Section VIII., a goal of training presentation models is to achieve the highest possible HLA coverage of each individual of a population, and therefore to HLA coverage of the population such that the proportions of individuals of the population with higher individual HLA coverages are as high as possible.

VII.B. Presentation Information

[00235] FIG. 2A illustrates a method of obtaining presentation information, in accordance with an embodiment. The presentation information 165 includes two general categories of information: allele-interacting information and allele-noninteracting information. Allele-interacting information includes information that influence presentation of peptide sequences that are dependent on the type of MHC allele. Allele-noninteracting information includes information that influence presentation of peptide sequences that are independent on the type of MHC allele.

VII.B.1. Allele-interacting Information

[00236] Allele-interacting information primarily includes identified peptide sequences that are known to have been presented by one or more identified MHC molecules from humans, mice, etc. Notably, this may or may not include data obtained from tumor samples. The presented peptide sequences may be identified from cells that express a single MHC allele. In this case the presented peptide sequences are generally collected from single-allele cell lines that are engineered to express a predetermined MHC allele and that are subsequently exposed to synthetic protein. Peptides presented on the MHC allele are isolated by techniques such as acid-elution and identified through mass spectrometry. FIG. 2B shows an example of this, where the example peptide YEMFNDKSQRAPDDKMF, presented on the predetermined MHC allele HLA-DRB1*12:01, is isolated and identified through mass spectrometry. Since in this situation peptides are identified through cells engineered to express a single predetermined MHC protein, the direct association between a presented peptide and the MHC protein to which it was bound to is definitively known.

[00237] The presented peptide sequences may also be collected from cells that express multiple MHC alleles. Typically in humans, 6 different types of MHC-I and up to 12 different types of MHC-II molecules are expressed for a cell. Such presented peptide sequences may be identified from multiple-allele cell lines that are engineered to express

multiple predetermined MHC alleles. Such presented peptide sequences may also be identified from tissue samples, either from normal tissue samples or tumor tissue samples. In this case particularly, the MHC molecules can be immunoprecipitated from normal or tumor tissue. Peptides presented on the multiple MHC alleles can similarly be isolated by techniques such as acid-elution and identified through mass spectrometry. FIG. 2C shows an example of this, where the six example peptides, YEMFNDKSF, HROEIFSHDFJ, FJIEJFOESS, NEIOREIREI, JFKSIFEMMSJDSSUIFLKSJFIEIFJ, and KNFLENFIESOFI, are presented on identified class I MHC alleles HLA-A*01:01, HLA-A*02:01, HLA-B*07:02, HLA-B*08:01, and class II MHC alleles HLA-DRB1*10:01, HLA-DRB1:11:01and are isolated and identified through mass spectrometry. In contrast to single-allele cell lines, the direct association between a presented peptide and the MHC protein to which it was bound to may be unknown since the bound peptides are isolated from the MHC molecules before being identified.

[00238] Allele-interacting information can also include mass spectrometry ion current which depends on both the concentration of peptide-MHC molecule complexes, and the ionization efficiency of peptides. The ionization efficiency varies from peptide to peptide in a sequence-dependent manner. Generally, ionization efficiency varies from peptide to peptide over approximately two orders of magnitude, while the concentration of peptide-MHC complexes varies over a larger range than that.

[00239] Allele-interacting information can also include measurements or predictions of binding affinity between a given MHC allele and a given peptide. (72, 73, 74) One or more affinity models can generate such predictions. For example, going back to the example shown in FIG. 1D, presentation information 165 may include a binding affinity prediction of 1000nM between the peptide YEMFNDKSF and the class I allele HLA-A*01:01. Few peptides with IC50 > 1000nm are presented by the MHC, and lower IC50 values increase the probability of presentation. Presentation information 165 may include a binding affinity prediction between the peptide KNFLENFIESOFI and the class II allele HLA-DRB1:11:01.

[00240] Allele-interacting information can also include measurements or predictions of stability of the MHC complex. One or more stability models that can generate such predictions. More stable peptide-MHC complexes (i.e., complexes with longer half-lives) are more likely to be presented at high copy number on tumor cells and on antigen-presenting cells that encounter vaccine antigen. For example, going back to the example shown in FIG. 2C, presentation information 165 may include a stability prediction of a half-life of 1h for the

class I molecule HLA-A*01:01. Presentation information 165 may also include a stability prediction of a half-life for the class II molecule HLA-DRB1:11:01.

[00241] Allele-interacting information can also include the measured or predicted rate of the formation reaction for the peptide-MHC complex. Complexes that form at a higher rate are more likely to be presented on the cell surface at high concentration.

[00242] Allele-interacting information can also include the sequence and length of the peptide. MHC class I molecules typically prefer to present peptides with lengths between 8 and 15 peptides. 60-80% of presented peptides have length 9. MHC class II molecules typically prefer to present peptides with lengths between 6-30 peptides.

[00243] Allele-interacting information can also include the presence of kinase sequence motifs on the neoantigen encoded peptide, and the absence or presence of specific post-translational modifications on the neoantigen encoded peptide. The presence of kinase motifs affects the probability of post-translational modification, which may enhance or interfere with MHC binding.

[00244] Allele-interacting information can also include the expression or activity levels of proteins involved in the process of post-translational modification, e.g., kinases (as measured or predicted from RNA seq, mass spectrometry, or other methods).

[00245] Allele-interacting information can also include the probability of presentation of peptides with similar sequence in cells from other individuals expressing the particular MHC allele as assessed by mass-spectrometry proteomics or other means.

[00246] Allele-interacting information can also include the expression levels of the particular MHC allele in the individual in question (e.g. as measured by RNA-seq or mass spectrometry). Peptides that bind most strongly to an MHC allele that is expressed at high levels are more likely to be presented than peptides that bind most strongly to an MHC allele that is expressed at a low level.

[00247] Allele-interacting information can also include the overall neoantigen encoded peptide-sequence-independent probability of presentation by the particular MHC allele in other individuals who express the particular MHC allele.

[00248] Allele-interacting information can also include the overall peptide-sequence-independent probability of presentation by MHC alleles in the same family of molecules (e.g., HLA-A, HLA-B, HLA-C, HLA-DQ, HLA-DR, HLA-DP) in other individuals. For example, HLA-C molecules are typically expressed at lower levels than HLA-A or HLA-B molecules, and consequently, presentation of a peptide by HLA-C is a priori less probable

than presentation by HLA-A or HLA-B. For another example, HLA-DP is typically expressed at lower levels than HLA-DR or HLA-DQ; consequently, presentation of a peptide by HLA-DP is a prior less probable than presentation by HLA-DR or HLA-DQ.

[00249] Allele-interacting information can also include the protein sequence of the particular MHC allele.

[00250] Any MHC allele-noninteracting information listed in the below section can also be modeled as an MHC allele-interacting information.

VII.B.2. Allele-noninteracting Information

[00251] Allele-noninteracting information can include C-terminal sequences flanking the neoantigen encoded peptide within its source protein sequence. For MHC-I, C-terminal flanking sequences may impact proteasomal processing of peptides. However, the C-terminal flanking sequence is cleaved from the peptide by the proteasome before the peptide is transported to the endoplasmic reticulum and encounters MHC alleles on the surfaces of cells. Consequently, MHC molecules receive no information about the C-terminal flanking sequence, and thus, the effect of the C-terminal flanking sequence cannot vary depending on MHC allele type. For example, going back to the example shown in FIG. 2C, presentation information 165 may include the C-terminal flanking sequence FOEIFNDKSLDKFJI of the presented peptide FJIEJFOESS identified from the source protein of the peptide.

[00252] Allele-noninteracting information can also include mRNA quantification measurements. For example, mRNA quantification data can be obtained for the same samples that provide the mass spectrometry training data. As later described, RNA expression was identified to be a strong predictor of peptide presentation. In one embodiment, the mRNA quantification measurements are identified from software tool RSEM. Detailed implementation of the RSEM software tool can be found at Bo Li and Colin N. Dewey. *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome*. BMC Bioinformatics, 12:323, August 2011. In one embodiment, the mRNA quantification is measured in units of fragments per kilobase of transcript per Million mapped reads (FPKM).

[00253] Allele-noninteracting information can also include the N-terminal sequences flanking the peptide within its source protein sequence.

[00254] Allele-noninteracting information can also include the source gene of the peptide sequence. The source gene may be defined as the Ensembl protein family of the peptide sequence. In other examples, the source gene may be defined as the source DNA or the

source RNA of the peptide sequence. The source gene can, for example, be represented as a string of nucleotides that encode for a protein, or alternatively be more categorically represented based on a named set of known DNA or RNA sequences that are known to encode specific proteins. In another example, allele-noninteracting information can also include the source transcript or isoform or set of potential source transcripts or isoforms of the peptide sequence drawn from a database such as Ensembl or RefSeq.

[00255] Allele-noninteracting information can also include the tissue type, cell type or tumor type of cells of origin of the peptide sequence.

[00256] Allele-noninteracting information can also include the presence of protease cleavage motifs in the peptide, optionally weighted according to the expression of corresponding proteases in the tumor cells (as measured by RNA-seq or mass spectrometry). Peptides that contain protease cleavage motifs are less likely to be presented, because they will be more readily degraded by proteases, and will therefore be less stable within the cell.

[00257] Allele-noninteracting information can also include the turnover rate of the source protein as measured in the appropriate cell type. Faster turnover rate (i.e., lower half-life) increases the probability of presentation; however, the predictive power of this feature is low if measured in a dissimilar cell type.

[00258] Allele-noninteracting information can also include the length of the source protein, optionally considering the specific splice variants ("isoforms") most highly expressed in the tumor cells as measured by RNA-seq or proteome mass spectrometry, or as predicted from the annotation of germline or somatic splicing mutations detected in DNA or RNA sequence data.

[00259] Allele-noninteracting information can also include the level of expression of the proteasome, immunoproteasome, thymoproteasome, or other proteases in the tumor cells (which may be measured by RNA-seq, proteome mass spectrometry, or immunohistochemistry). Different proteasomes have different cleavage site preferences. More weight will be given to the cleavage preferences of each type of proteasome in proportion to its expression level.

[00260] Allele-noninteracting information can also include the expression of the source gene of the peptide (e.g., as measured by RNA-seq or mass spectrometry). Possible optimizations include adjusting the measured expression to account for the presence of stromal cells and tumor-infiltrating lymphocytes within the tumor sample. Peptides from

more highly expressed genes are more likely to be presented. Peptides from genes with undetectable levels of expression can be excluded from consideration.

[00261] Allele-noninteracting information can also include the probability that the source mRNA of the neoantigen encoded peptide will be subject to nonsense-mediated decay as predicted by a model of nonsense-mediated decay, for example, the model from Rivas et al, Science 2015.

[00262] Allele-noninteracting information can also include the typical tissue-specific expression of the source gene of the peptide during various stages of the cell cycle. Genes that are expressed at a low level overall (as measured by RNA-seq or mass spectrometry proteomics) but that are known to be expressed at a high level during specific stages of the cell cycle are likely to produce more presented peptides than genes that are stably expressed at very low levels.

[00263] Allele-noninteracting information can also include a comprehensive catalog of features of the source protein as given in e.g. uniProt or PDB http://www.rcsb.org/pdb/home/home.do. These features may include, among others: the secondary and tertiary structures of the protein, subcellular localization 11, Gene ontology (GO) terms. Specifically, this information may contain annotations that act at the level of the protein, e.g., 5' UTR length, and annotations that act at the level of specific residues, e.g., helix motif between residues 300 and 310. These features can also include turn motifs, sheet motifs, and disordered residues.

[00264] Allele-noninteracting information can also include features describing the properties of the domain of the source protein containing the peptide, for example: secondary or tertiary structure (e.g., alpha helix vs beta sheet); Alternative splicing.

[00265] Allele-noninteracting information can also include associations between a peptide sequence of the neoantigen and one or more k-mer blocks of a plurality of k-mer blocks of a source gene of the neoantigen (as present in the nucleotide sequencing data of the subject). During training of the presentation model, these associations between the peptide sequence of the neoantigen and the k-mer blocks of the nucleotide sequencing data of the neoantigen are input into the model, and are used in part by the model to learn model parameters that represent presence or absence of a presentation hotspot for the k-mer blocks associated with the training peptide sequences. Then, during use of the model subsequent to training, associations between a test peptide sequence and one or more k-mer blocks of a source gene of test peptide sequence are input into the model, and the parameters learned by the model

during training enable the presentation model to make more accurate predictions regarding the presentation likelihood of the test peptide sequence.

[00266] In general, the parameters of the model that represent presence or absence of a presentation hotspot for a k-mer block represent the residual propensity that the k-mer block will give rise to presented peptides, after controlling for all other variables (e.g., peptide sequence, RNA expression, amino acids commonly found in HLA-binding peptides, etc.). The parameters representing presence or absence of a presentation hotspot for a k-mer block may be a binary coefficient (e.g., 0 or 1), or an analog coefficient along a scale (e.g., between 0 and 1, inclusive). In either case, a greater coefficient (e.g., closer to 1 or 1) represents a greater likelihood that the k-mer block will give rise to presented peptides controlling for other factors, whereas lower coefficient (e.g., closer to 0 or 0) represents a lower likelihood that the k-mer block will give rise to presented peptides. For example, a k-mer block with a low hotspot coefficient might be a k-mer block from a gene with high RNA expression, with amino acids commonly found in HLA-binding peptides, where the source gene gives rise to lots of other presented peptides, but where presented peptides are rarely seen in the k-mer block. Since other sources of peptide presence may already be accounted for by other parameters (e.g., RNA expression on a k-mer block or larger basis, commonly found in HLA-binding peptides), these hotspot parameters provide new, separate information that does not "double count" information captured by other parameters.

[00267] Allele-noninteracting information can also include the probability of presentation of peptides from the source protein of the peptide in question in other individuals (after adjusting for the expression level of the source protein in those individuals and the influence of the different HLA types of those individuals).

[00268] Allele-noninteracting information can also include the probability that the peptide will not be detected or over-represented by mass spectrometry due to technical biases.

[00269] The expression of various gene modules/pathways as measured by a gene expression assay such as RNASeq, microarray(s), targeted panel(s) such as Nanostring, or single/multi- gene representatives of gene modules measured by assays such as RT-PCR (which need not contain the source protein of the peptide) that are informative about the state of the tumor cells, stroma, or tumor-infiltrating lymphocytes (TILs).

[00270] Allele-noninteracting information can also include the copy number of the source gene of the peptide in the tumor cells. For example, peptides from genes that are subject to homozygous deletion in tumor cells can be assigned a probability of presentation of zero.

[00271] Allele-noninteracting information can also include the probability that the peptide binds to the TAP or the measured or predicted binding affinity of the peptide to the TAP. Peptides that are more likely to bind to the TAP, or peptides that bind the TAP with higher affinity are more likely to be presented by MHC-I.

[00272] Allele-noninteracting information can also include the expression level of TAP in the tumor cells (which may be measured by RNA-seq, proteome mass spectrometry, immunohistochemistry). For MHC-I, higher TAP expression levels increase the probability of presentation of all peptides.

[00273] Allele-noninteracting information can also include the presence or absence of tumor mutations, including, but not limited to:

i. Driver mutations in known cancer driver genes such as EGFR, KRAS, ALK, RET, ROS1, TP53, CDKN2A, CDKN2B, NTRK1, NTRK2, NTRK3

ii. In genes encoding the proteins involved in the antigen presentation machinery (e.g., B2M, HLA-A, HLA-B, HLA-C, TAP-1, TAP-2, TAPBP, CALR, CNX, ERP57, HLA-DM, HLA-DMA, HLA-DMB, HLA-DO, HLA-DOA, HLA-DOBHLA-DP, HLA-DPA1, HLA-DPB1, HLA-DQ, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DR, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5 or any of the genes coding for components of the proteasome or immunoproteasome). Peptides whose presentation relies on a component of the antigen-presentation machinery that is subject to loss-of-function mutation in the tumor have reduced probability of presentation.

[00274] Presence or absence of functional germline polymorphisms, including, but not limited to:

i. In genes encoding the proteins involved in the antigen presentation machinery (e.g., B2M, HLA-A, HLA-B, HLA-C, TAP-1, TAP-2, TAPBP, CALR, CNX, ERP57, HLA-DM, HLA-DMA, HLA-DMB, HLA-DO, HLA-DOA, HLA-DOBHLA-DP, HLA-DPA1, HLA-DPB1, HLA-DQ, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DR, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5 or any of the genes coding for components of the proteasome or immunoproteasome)

[00275] Allele-noninteracting information can also include tumor type (e.g., NSCLC, melanoma).

[00276] Allele-noninteracting information can also include known functionality of HLA alleles, as reflected by, for instance HLA allele suffixes. For example, the N suffix in the

allele name HLA-A*24:09N indicates a null allele that is not expressed and is therefore unlikely to present epitopes; the full HLA allele suffix nomenclature is described at https://www.ebi.ac.uk/ipd/imgt/hla/nomenclature/suffixes.html.

[00277] Allele-noninteracting information can also include clinical tumor subtype (e.g., squamous lung cancer vs. non-squamous).

[00278] Allele-noninteracting information can also include smoking history.

[00279] Allele-noninteracting information can also include history of sunburn, sun exposure, or exposure to other mutagens.

[00280] Allele-noninteracting information can also include the typical expression of the source gene of the peptide in the relevant tumor type or clinical subtype, optionally stratified by driver mutation. Genes that are typically expressed at high levels in the relevant tumor type are more likely to be presented.

[00281] Allele-noninteracting information can also include the frequency of the mutation in all tumors, or in tumors of the same type, or in tumors from individuals with at least one shared MHC allele, or in tumors of the same type in individuals with at least one shared MHC allele.

[00282] In the case of a mutated tumor-specific peptide, the list of features used to predict a probability of presentation may also include the annotation of the mutation (e.g., missense, read-through, frameshift, fusion, etc.) or whether the mutation is predicted to result in nonsense-mediated decay (NMD). For example, peptides from protein segments that are not translated in tumor cells due to homozygous early-stop mutations can be assigned a probability of presentation of zero. NMD results in decreased mRNA translation, which decreases the probability of presentation.

VII.C. Presentation Identification System

[00283] FIG. 3 is a high-level block diagram illustrating the computer logic components of the presentation identification system 160, according to one embodiment. In this example embodiment, the presentation identification system 160 includes a data management module 312, an encoding module 314, a training module 316, and a prediction module 320. The presentation identification system 160 is also comprised of a training data store 170 and a presentation models store 175. Some embodiments of the model management system 160 have different modules than those described here. Similarly, the functions can be distributed among the modules in a different manner than is described here.

VII.C.1. Data Management Module

[00284] The data management module 312 generates sets of training data 170 from the presentation information 165. Each set of training data contains a plurality of data instances, in which each data instance i contains a set of independent variables z^i that include at least a presented or non-presented peptide sequence p^i, one or more associated MHC alleles a^i associated with the peptide sequence p^i and/or one or more MHC allele sequences d^i associated with the peptide sequence p^i, and a dependent variable y^i that represents information that the presentation identification system 160 is interested in predicting for new values of independent variables.

[00285] In one particular implementation referred throughout the remainder of the specification, the dependent variable y^i is a binary label indicating whether peptide p^i was presented by the one or more associated MHC alleles a^i and/or by one or more MHC alleles associated with the one or more MHC allele sequences d^i. However, it is appreciated that in other implementations, the dependent variable y^i can represent any other kind of information that the presentation identification system 160 is interested in predicting dependent on the independent variables z^i. For example, in another implementation, the dependent variable y^i may also be a numerical value indicating the mass spectrometry ion current identified for the data instance.

[00286] The peptide sequence p^i for data instance i is a sequence of k_i amino acids, in which k_i may vary between data instances i within a range. For example, that range may be 8-15 for MHC class I or 6-30 for MHC class II. In one specific implementation of system 160, all peptide sequences p^i in a training data set may have the same length, e.g. 9. The number of amino acids in a peptide sequence may vary depending on the type of MHC alleles (e.g., MHC alleles in humans, etc.). The MHC alleles a^i for data instance i indicate which MHC alleles were present in association with the corresponding peptide sequence p^i. Similarity, in some embodiments, the MHC allele sequences d^i for data instance i indicate which MHC allele sequences were present in association with the corresponding peptide sequence p^i.

[00287] The data management module 312 may also include additional allele-interacting variables, such as binding affinity b^i and stability s^i predictions in conjunction with the peptide sequences p^i and associated MHC alleles a^i contained in the training data 170. For example, the training data 170 may contain binding affinity predictions b^i between a peptide p^i and each of the associated MHC molecules indicated in a^i. As another example, the

training data 170 may contain stability predictions s^i for each of the MHC alleles indicated in a^i.

[00288] The data management module 312 may also include allele-noninteracting variables w^i, such as C-terminal flanking sequences and mRNA quantification measurements in conjunction with the peptide sequences p^i.

[00289] The data management module 312 also identifies peptide sequences that are not presented by MHC alleles to generate the training data 170. Generally, this involves identifying the "longer" sequences of source protein that include presented peptide sequences prior to presentation. When the presentation information contains engineered cell lines, the data management module 312 identifies a series of peptide sequences in the synthetic protein to which the cells were exposed to that were not presented on MHC alleles of the cells. When the presentation information contains tissue samples, the data management module 312 identifies source proteins from which presented peptide sequences originated from, and identifies a series of peptide sequences in the source protein that were not presented on MHC alleles of the tissue sample cells.

[00290] The data management module 312 may also artificially generate peptides with random sequences of amino acids and identify the generated sequences as peptides not presented on MHC alleles. This can be accomplished by randomly generating peptide sequences allows the data management module 312 to easily generate large amounts of synthetic data for peptides not presented on MHC alleles. Since in reality, a small percentage of peptide sequences are presented by MHC alleles, the synthetically generated peptide sequences are highly likely not to have been presented by MHC alleles even if they were included in proteins processed by cells.

[00291] FIG. 4 illustrates an example set of training data 170A, according to one embodiment. Specifically, the first 3 data instances in the training data 170A indicate peptide presentation information from a single-allele cell line involving the allele HLA-C*01:03 and 3 peptide sequences QCEIOWAREFLKEIGJ, FIEUHFWI, and FEWRHRJTRUJR. Note that in alternative embodiments of the training data 170A, the HLA allele type may be replaced by the HLA allele sequence. For instance, the allele type HLA-C*1:03 may be replaced by the amino acid sequence for the allele HLA-C*1:03. The fourth data instance in the training data 170A indicates peptide information from a multiple-allele cell line involving the alleles HLA-B*07:02, HLA-C*01:03, HLA-A*01:01and a peptide sequence QIEJOEIJE. The first data instance indicates that peptide sequence QCEIOWARE was not presented by the allele

HLA-DRB3:01:01. As discussed in the prior two paragraphs, the negatively-labeled peptide sequences may be randomly generated by the data management module 312 or identified from source protein of presented peptides. The training data 170A also includes a binding affinity prediction of 1000nM and a stability prediction of a half-life of 1h for the peptide sequence-allele pair. The training data 170A also includes allele-noninteracting variables, such as the C-terminal flanking sequence of the peptide FJELFISBOSJFIE, and a mRNA quantification measurement of 10^2 TPM. The fourth data instance indicates that peptide sequence QIEJOEIJE was presented by one of the alleles HLA-B*07:02, HLA-C*01:03, or HLA-A*01:01. The training data 170A also includes binding affinity predictions and stability predictions for each of the alleles, as well as the C-terminal flanking sequence of the peptide and the mRNA quantification measurement for the peptide. In further embodiments, the training data 170A may also include additional allele-noninteracting variables such as peptide families of the presented peptides.

VII.C.2. Encoding Module

[00292] The encoding module 314 encodes information contained in the training data 170 into a numerical representation that can be used to generate the one or more presentation models. In one implementation, the encoding module 314 one-hot encodes sequences (e.g., peptide sequences and/or C-terminal flanking sequences and/or MHC allele sequences) over a predetermined 20-letter amino acid alphabet. Specifically, a peptide sequence p^i with k_i amino acids is represented as a row vector of $20 \cdot k_i$ elements, where a single element among $p^i_{20 \cdot (j-1)+1}$, $p^i_{20 \cdot (j-1)+2}$, ..., $p^i_{20 \cdot j}$ that corresponds to the alphabet of the amino acid at the *j-th* position of the peptide sequence has a value of 1. Otherwise, the remaining elements have a value of 0. As an example, for a given alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, the peptide sequence EAF of 3 amino acids for data instance *i* may be represented by the row vector of 60 elements p^i=[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

170. Thus, when the peptide sequence with the greatest length has k_{max} amino acids, the encoding module 314 numerically represents each sequence as a row vector of $(20+1)\cdot k_{max}$ elements. As an example, for the extended alphabet {PAD, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} and a maximum amino acid length of $k_{max}=5$, the same example peptide sequence EAF of 3 amino acids may be represented by the row vector of 105 elements $\boldsymbol{p}^i$=[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

[00297] The encoding module 314 may represent a pair of allele-interacting variables x_h^i for peptide p_i and an associated MHC allele h as a row vector in which numerical representations of allele-interacting variables are concatenated one after the other. For example, the encoding module 314 may represent x_h^i as a row vector equal to $[p^i]$, $[p^i\ b_h^i]$, $[p^i\ s_h^i]$, or $[p^i\ b_h^i\ s_h^i]$, where b_h^i is the binding affinity prediction for peptide p_i and associated MHC allele h, and similarly for s_h^i for stability. Alternatively, one or more combination of allele-interacting variables may be stored individually (e.g., as individual vectors or matrices).

[00298] In one instance, the encoding module 314 represents binding affinity information by incorporating measured or predicted values for binding affinity in the allele-interacting variables x_h^i.

[00299] In one instance, the encoding module 314 represents binding stability information by incorporating measured or predicted values for binding stability in the allele-interacting variables x_h^i,

[00300] In one instance, the encoding module 314 represents binding on-rate information by incorporating measured or predicted values for binding on-rate in the allele-interacting variables x_h^i.

[00301] In one instance, for peptides presented by class I MHC molecules, the encoding module 314 represents peptide length as a vector T_k=[$\mathbb{1}(L_k$=8) $\mathbb{1}(L_k$=9) $\mathbb{1}(L_k$=10) $\mathbb{1}(L_k$=11) $\mathbb{1}(L_k$=12) $\mathbb{1}(L_k$=13) $\mathbb{1}(L_k$=14) $\mathbb{1}(L_k$=15)] where $\mathbb{1}$ is the indicator function, and L_k denotes the length of peptide p_k. The vector T_k can be included in the allele-interacting variables x_h^i. In another instance, for peptides presented by class II MHC molecules, the encoding module 314 represents peptide length as a vector T_k=[$\mathbb{1}(L_k$=6) $\mathbb{1}(L_k$=7) $\mathbb{1}(L_k$=8) $\mathbb{1}(L_k$=9) $\mathbb{1}(L_k$=10) $\mathbb{1}(L_k$=11) $\mathbb{1}(L_k$=12) $\mathbb{1}(L_k$=13) $\mathbb{1}(L_k$=14) $\mathbb{1}(L_k$=15) $\mathbb{1}(L_k$=16) $\mathbb{1}(L_k$=17) $\mathbb{1}(L_k$=18) $\mathbb{1}(L_k$=19) $\mathbb{1}(L_k$=20) $\mathbb{1}(L_k$=21) $\mathbb{1}(L_k$=22) $\mathbb{1}(L_k$=23) $\mathbb{1}(L_k$=24) $\mathbb{1}(L_k$=25) $\mathbb{1}(L_k$=26) $\mathbb{1}(L_k$=27) $\mathbb{1}(L_k$=28) $\mathbb{1}(L_k$=29) $\mathbb{1}(L_k$=30)] where $\mathbb{1}$ is the indicator function, and L_k denotes the length of peptide p_k. The vector T_k can be included in the allele-interacting variables x_h^i.

[00302] In one instance, the encoding module 314 represents RNA expression information of MHC alleles by incorporating RNA-seq based expression levels of MHC alleles in the allele-interacting variables x_h^i.

[00303] Similarly, the encoding module 314 may represent the allele-noninteracting variables w^i as a row vector in which numerical representations of allele-noninteracting variables are concatenated one after the other. For example, w^i may be a row vector equal to

$[c^i]$ or $[c^i\ m^i\ w^i]$ in which w^i is a row vector representing any other allele-noninteracting variables in addition to the C-terminal flanking sequence of peptide p^i and the mRNA quantification measurement m^i associated with the peptide. Alternatively, one or more combination of allele-noninteracting variables may be stored individually (e.g., as individual vectors or matrices).

[00304] In one instance, the encoding module 314 represents turnover rate of source protein for a peptide sequence by incorporating the turnover rate or half-life in the allele-noninteracting variables w^i.

[00305] In one instance, the encoding module 314 represents length of source protein or isoform by incorporating the protein length in the allele-noninteracting variables w^i.

[00306] In one instance, the encoding module 314 represents activation of immunoproteasome by incorporating the mean expression of the immunoproteasome-specific proteasome subunits including the $\beta 1_i$, $\beta 2_i$, $\beta 5_i$ subunits in the allele-noninteracting variables w^i.

[00307] In one instance, the encoding module 314 represents the RNA-seq abundance of the source protein of the peptide or gene or transcript of a peptide (quantified in units of FPKM, TPM by techniques such as RSEM) can be incorporating the abundance of the source protein in the allele-noninteracting variables w^i.

[00308] In one instance, the encoding module 314 represents the probability that the transcript of origin of a peptide will undergo nonsense-mediated decay (NMD) as estimated by the model in, for example, Rivas et. al. *Science*, 2015 by incorporating this probability in the allele-noninteracting variables w^i.

[00309] In one instance, the encoding module 314 represents the activation status of a gene module or pathway assessed via RNA-seq by, for example, quantifying expression of the genes in the pathway in units of TPM using e.g., RSEM for each of the genes in the pathway then computing a summary statistics, e.g., the mean, across genes in the pathway. The mean can be incorporated in the allele-noninteracting variables w^i.

[00310] In one instance, the encoding module 314 represents the copy number of the source gene by incorporating the copy number in the allele-noninteracting variables w^i.

[00311] In one instance, the encoding module 314 represents the TAP binding affinity by including the measured or predicted TAP binding affinity (e.g., in nanomolar units) in the allele-noninteracting variables w^i.

[00312] In one instance, the encoding module 314 represents TAP expression levels by including TAP expression levels measured by RNA-seq (and quantified in units of TPM by e.g., RSEM) in the allele-noninteracting variables w^j.

[00313] In one instance, the encoding module 314 represents tumor mutations as a vector of indicator variables (i.e., $d^k = 1$ if peptide p^k comes from a sample with a KRAS G12D mutation and 0 otherwise) in the allele-noninteracting variables w^j.

[00314] In one instance, the encoding module 314 represents germline polymorphisms in antigen presentation genes as a vector of indicator variables (i.e., $d^k = 1$ if peptide p^k comes from a sample with a specific germline polymorphism in the TAP). These indicator variables can be included in the allele-noninteracting variables w^j.

[00315] In one instance, the encoding module 314 represents tumor type as a length-one one-hot encoded vector over the alphabet of tumor types (e.g., NSCLC, melanoma, colorectal cancer, etc). These one-hot-encoded variables can be included in the allele-noninteracting variables w^j.

[00316] In one instance, the encoding module 314 represents MHC allele suffixes by treating 4-digit HLA alleles with different suffixes. For example, HLA-A*24:09N is considered a different allele from HLA-A*24:09 for the purpose of the model. Alternatively, the probability of presentation by an N-suffixed MHC allele can be set to zero for all peptides, because HLA alleles ending in the N suffix are not expressed.

[00317] In one instance, the encoding module 314 represents tumor subtype as a length-one one-hot encoded vector over the alphabet of tumor subtypes (e.g., lung adenocarcinoma, lung squamous cell carcinoma, etc). These one-hot encoded variables can be included in the allele-noninteracting variables w^j.

[00318] In one instance, the encoding module 314 represents smoking history as a binary indicator variable ($d^k = 1$ if the patient has a smoking history, and 0 otherwise), that can be included in the allele-noninteracting variables w^j. Alternatively, smoking history can be encoded as a length-one one-hot encoded variable over an alphabet of smoking severity. For example, smoking status can be rated on a 1-5 scale, where 1 indicates nonsmokers, and 5 indicates current heavy smokers. Because smoking history is primarily relevant to lung tumors, when training a model on multiple tumor types, this variable can also be defined to be equal to 1 if the patient has a history of smoking and the tumor type is lung tumors and zero otherwise.

[00319] In one instance, the encoding module 314 represents sunburn history as a binary indicator variable ($d^k = 1$ if the patient has a history of severe sunburn, and 0 otherwise), which can be included in the allele-noninteracting variables w^j. Because severe sunburn is primarily relevant to melanomas, when training a model on multiple tumor types, this variable can also be defined to be equal to 1 if the patient has a history of severe sunburn and the tumor type is melanoma and zero otherwise.

[00320] In one instance, the encoding module 314 represents distribution of expression levels of a particular gene or transcript for each gene or transcript in the human genome as summary statistics (e,g., mean, median) of distribution of expression levels by using reference databases such as TCGA. Specifically, for a peptide p^k in a sample with tumor type melanoma, not only the measured gene or transcript expression level of the gene or transcript of origin of peptide p^k in the allele-noninteracting variables w^j, but also the mean and/or median gene or transcript expression of the gene or transcript of origin of peptide p^k in melanomas as measured by TCGA can be included.

[00321] In one instance, the encoding module 314 represents mutation type as a length-one one-hot-encoded variable over the alphabet of mutation types (e.g., missense, frameshift, NMD-inducing, etc). These onehot-encoded variables can be included in the allele-noninteracting variables w^j.

[00322] In one instance, the encoding module 314 represents protein-level features of protein as the value of the annotation (e.g., 5' UTR length) of the source protein in the allele-noninteracting variables w^j. In another instance, the encoding module 314 represents residue-level annotations of the source protein for peptide p^j by including an indicator variable, that is equal to 1 if peptide p^j overlaps with a helix motif and 0 otherwise, or that is equal to 1 if peptide p^j is completely contained with within a helix motif in the allele-noninteracting variables w^j. In another instance, a feature representing proportion of residues in peptide p^j that are contained within a helix motif annotation can be included in the allele-noninteracting variables w^j.

[00323] In one instance, the encoding module 314 represents type of proteins or isoforms in the human proteome as an indicator vector o^k that has a length equal to the number of proteins or isoforms in the human proteome, and the corresponding element o^k_i is 1 if peptide p^k comes from protein i and 0 otherwise.

[00324] In one instance, the encoding module 314 represents the source gene G=gene(p^i) of peptide p^i as a categorical variable with L possible categories, where L denotes the upper limit of the number of indexed source genes 1, 2, ..., L.

[00325] In one instance, the encoding module 314 represents the tissue type, cell type, tumor type, or tumor histology type T=tissue(p^i) of peptide p^i as a categorical variable with M possible categories, where M denotes the upper limit of the number of indexed types 1, 2, ..., M. Types of tissue can include, for example, lung tissue, cardiac tissue, intestine tissue, nerve tissue, and the like. Types of cells can include dendritic cells, macrophages, CD4 T cells, and the like. Types of tumors can include lung adenocarcinoma, lung squamous cell carcinoma, melanoma, non-Hodgkin lymphoma, and the like.

[00326] The encoding module 314 may also represent the overall set of variables z^i for peptide p^i and an associated MHC allele h as a row vector in which numerical representations of the allele-interacting variables x^i and the allele-noninteracting variables w^i are concatenated one after the other. For example, the encoding module 314 may represent $z_h{}^i$ as a row vector equal to $[x_h{}^i\ w^i]$ or $[w_i\ x_h{}^i]$.

VIII. Training Module

[00327] The training module 316 constructs one or more presentation models that generate likelihoods of whether peptide sequences will be presented by MHC alleles associated with the peptide sequences. Specifically, given a peptide sequence p^k and a set of MHC alleles a^k and/or MHC allele sequences d^k associated with the peptide sequence p^k, each presentation model generates an estimate u_k indicating a likelihood that the peptide sequence p^k will be presented by one or more of the associated MHC alleles a^k.

VIII.A. Overview

[00328] The training module 316 constructs the one more presentation models based on the training data sets stored in store 170 generated from the presentation information stored in 165. Generally, regardless of the specific type of presentation model, all of the presentation models capture the dependence between independent variables and dependent variables in the training data 170 such that a loss function is minimized. Specifically, the loss function $\ell(y_{i\in S}, u_{i\in S}, \theta)$ represents discrepancies between values of dependent variables $y_{i\in S}$ for one or more data instances S in the training data 170 and the estimated likelihoods $u_{i\in S}$ for the data instances S generated by the presentation model. In one particular implementation referred

throughout the remainder of the specification, the loss function ($y_{i\in S}$, $u_{i\in S}$, θ) is the negative log likelihood function given by equation (1a) as follows:

$$\ell(y_{i\in S}, u_{i\in S}; \theta) = \sum_{i\in S} (y_i \log u_i + (1 - y_i) \log(1 - u_i)). \quad (1a)$$

However, in practice, another loss function may be used. For example, when predictions are made for the mass spectrometry ion current, the loss function is the mean squared loss given by equation 1b as follows:

$$\ell(y_{i\in S}, u_{i\in S}; \theta) = \sum_{i\in S} (\|y_i - u_i\|_2^2). \quad (1b)$$

[00329] The presentation model may be a parametric model in which one or more parameters θ mathematically specify the dependence between the independent variables and dependent variables. Typically, various parameters of parametric-type presentation models that minimize the loss function ($y_{i\in S}$, $u_{i\in S}$, θ) are determined through gradient-based numerical optimization algorithms, such as batch gradient algorithms, stochastic gradient algorithms, and the like. Alternatively, the presentation model may be a non-parametric model in which the model structure is determined from the training data 170 and is not strictly based on a fixed set of parameters.

VIII.B. Per-Allele Models

[00330] The training module 316 may construct the presentation models to predict presentation likelihoods of peptides on a per-allele basis. In this case, the training module 316 may train the presentation models based on data instances S in the training data 170 generated from cells expressing single MHC alleles.

[00331] In one implementation, the training module 316 models the estimated presentation likelihood u_k for peptide $\boldsymbol{p}^k$ for a specific allele h by:

$$u_k^h = \Pr(\boldsymbol{p}^k \text{ presented; MHC allele } h) = f\left(g_h(\boldsymbol{x}_h^k; \boldsymbol{\theta}_h)\right), \quad (2)$$

where $\boldsymbol{x}_h^k$ denotes the encoded allele-interacting variables for peptide $\boldsymbol{p}^k$ and corresponding MHC allele h, $f(\cdot)$ is any function, and is herein throughout is referred to as a transformation function for convenience of description. Further, $g_h(\cdot)$ is any function, is herein throughout referred to as a dependency function for convenience of description, and generates dependency scores for the allele-interacting variables $\boldsymbol{x}_h^k$ based on a set of parameters $\boldsymbol{\theta}_h$ determined for MHC allele h. The values for the set of parameters $\boldsymbol{\theta}_h$ for each MHC allele h can be determined by minimizing the loss function with respect to $\boldsymbol{\theta}_h$, where i is each instance in the subset S of training data 170 generated from cells expressing the single MHC allele h.

[00332] The output of the dependency function $g_h(x_h^k;\theta_h)$ represents a dependency score for the MHC allele h indicating whether the MHC allele h will present the corresponding neoantigen based on at least the allele interacting features x_h^k, and in particular, based on positions of amino acids of the peptide sequence of peptide p^k. For example, the dependency score for the MHC allele h may have a high value if the MHC allele h is likely to present the peptide p^k, and may have a low value if presentation is not likely. The transformation function $f(\cdot)$ transforms the input, and more specifically, transforms the dependency score generated by $g_h(x_h^k;\theta_h)$ in this case, to an appropriate value to indicate the likelihood that the peptide p^k will be presented by an MHC allele.

[00333] In one particular implementation referred throughout the remainder of the specification, $f(\cdot)$ is a function having the range within [0, 1] for an appropriate domain range. In one example, $f(\cdot)$ is the expit function given by:

$$f(z) = \frac{\exp(z)}{1+\exp(z)}. \tag{3}$$

As another example, $f(\cdot)$ can also be the hyperbolic tangent function given by:

$$f(z) = \tanh(z) \tag{4}$$

when the values for the domain z is equal to or greater than 0. Alternatively, when predictions are made for the mass spectrometry ion current that have values outside the range [0, 1], $f(\cdot)$ can be any function such as the identity function, the exponential function, the log function, and the like.

[00334] Thus, the per-allele likelihood that a peptide sequence p^k will be presented by a MHC allele h can be generated by applying the dependency function $g_h(\cdot)$ for the MHC allele h to the encoded version of the peptide sequence p^k to generate the corresponding dependency score. The dependency score may be transformed by the transformation function $f(\cdot)$ to generate a per-allele likelihood that the peptide sequence p^k will be presented by the MHC allele h.

VIII.B.1 Dependency Functions for Allele Interacting Variables

[00335] In one particular implementation referred throughout the specification, the dependency function $g_h(\cdot)$ is an affine function given by:

$$g_h(x_h^i;\theta_h) = x_h^i \cdot \theta_h. \tag{5}$$

that linearly combines each allele-interacting variable in x_h^k with a corresponding parameter in the set of parameters θ_h determined for the associated MHC allele h.

[00336] In another particular implementation referred throughout the specification, the dependency function $g_h(\cdot)$ is a network function given by:

$$g_h(x_h^i; \theta_h) = NN_h(x_h^i; \theta_h). \quad (6)$$

represented by a network model $NN_h(\cdot)$ having a series of nodes arranged in one or more layers. A node may be connected to other nodes through connections each having an associated parameter in the set of parameters θ_h. A value at one particular node may be represented as a sum of the values of nodes connected to the particular node weighted by the associated parameter mapped by an activation function associated with the particular node. In contrast to the affine function, network models are advantageous because the presentation model can incorporate non-linearity and process data having different lengths of amino acid sequences. Specifically, through non-linear modeling, network models can capture interaction between amino acids at different positions in a peptide sequence and how this interaction affects peptide presentation.

[00337] In general, network models $NN_h(\cdot)$ may be structured as feed-forward networks, such as artificial neural networks (ANN), convolutional neural networks (CNN), deep neural networks (DNN), and/or recurrent networks, such as long short-term memory networks (LSTM), bi-directional recurrent networks, deep bi-directional recurrent networks, and the like.

[00338] In one instance referred throughout the remainder of the specification, each MHC allele in $h=1,2,\ldots, m$ is associated with a separate network model, and $NN_h(\cdot)$ denotes the output(s) from a network model associated with MHC allele h.

[00339] FIG. 5 illustrates an example network model $NN_3(\cdot)$ in association with an arbitrary MHC allele $h=3$. As shown in FIG. 5, the network model $NN_3(\cdot)$ for MHC allele $h=3$ includes three input nodes at layer $l=1$, four nodes at layer $l=2$, two nodes at layer $l=3$, and one output node at layer $l=4$. The network model $NN_3(\cdot)$ is associated with a set of ten parameters $\theta_3(1)$, $\theta_3(2)$, ..., $\theta_3(10)$. The network model $NN_3(\cdot)$ receives input values (individual data instances including encoded polypeptide sequence data and any other training data used) for three allele-interacting variables $x_3^k(1)$, $x_3^k(2)$, and $x_3^k(3)$ for MHC allele $h=3$ and outputs the value $NN_3(x_3^k)$. The network function may also include one or more network models each taking different allele interacting variables as input.

[00340] In another instance, the identified MHC alleles $h=1, 2, \ldots, m$ are associated with a single network model $NN_H(\cdot)$, and $NN_h(\cdot)$ denotes one or more outputs of the single network model associated with MHC allele h. In such an instance, the set of parameters θ_h may

correspond to a set of parameters for the single network model, and thus, the set of parameters θ_h may be shared by all MHC alleles.

[00341] FIG. 6 illustrates an example network model $NN_H(\cdot)$ shared by MHC alleles $h=1,2, \ldots,m$. As shown in FIG. 6, the network model $NN_H(\cdot)$includes m output nodes each corresponding to an MHC allele. The network model $NN_3(\cdot)$ receives the allele-interacting variables x_3^k for MHC allele $h=3$ and outputs m values including the value $NN_3(x_3^k)$ corresponding to the MHC allele $h=3$.

[00342] In yet another instance, the dependency function $g_h(\cdot)$ can be expressed as:

$$g_h(x_h^k; \theta_h) = g'_h(x_h^k; \theta'_h) + \theta_h^0$$

where $g'_h(x_h^k;\theta'_h)$ is the affine function with a set of parameters θ'_h, the network function, or the like, with a bias parameter θ_h^0 in the set of parameters for allele interacting variables for the MHC allele that represents a baseline probability of presentation for the MHC allele h.

[00343] In another implementation, the bias parameter θ_h^0 may be shared according to the gene family of the MHC allele h. That is, the bias parameter θ_h^0 for MHC allele h may be equal to $\theta_{gene(h)}^0$, where $gene(h)$ is the gene family of MHC allele h. For example, class I MHC alleles HLA-A*02:01, HLA-A*02:02, and HLA-A*02:03 may be assigned to the gene family of "HLA-A," and the bias parameter θ_h^0 for each of these MHC alleles may be shared. As another example, class II MHC alleles HLA-DRB1:10:01, HLA-DRB1:11:01, and HLA-DRB3:01:01 may be assigned to the gene family of "HLA-DRB," and the bias parameter θ_h^0 for each of these MHC alleles may be shared.

[00344] Returning to equation (2), as an example, the likelihood that peptide p^k will be presented by MHC allele $h=3$, among $m=4$ different identified MHC alleles using the affine dependency function $g_h(\cdot)$, can be generated by:

$$u_k^3 = f(x_3^k \cdot \theta_3),$$

where x_3^k are the identified allele-interacting variables for MHC allele $h=3$, and θ_3 are the set of parameters determined for MHC allele $h=3$ through loss function minimization.

[00345] As another example, the likelihood that peptide p^k will be presented by MHC allele $h=3$, among $m=4$ different identified MHC alleles using separate network transformation functions $g_h(\cdot)$, can be generated by:

$$u_k^3 = f(NN_3(x_3^k; \theta_3)),$$

where x_3^k are the identified allele-interacting variables for MHC allele $h=3$, and θ_3 are the set of parameters determined for the network model $NN_3(\cdot)$ associated with MHC allele $h=3$.

[00346] FIG. 7 illustrates generating a presentation likelihood for peptide p^k in association with MHC allele $h=3$ using an example network model $NN_3(\cdot)$. As shown in FIG. 7, the network model $NN_3(\cdot)$ receives the allele-interacting variables x_3^k for MHC allele $h=3$ and generates the output $NN_3(x_3^k)$. The output is mapped by function $f(\cdot)$ to generate the estimated presentation likelihood u_k.

VIII.B.2. Per-Allele with Allele-Noninteracting Variables

[00347] In one implementation, the training module 316 incorporates allele-noninteracting variables and models the estimated presentation likelihood u_k for peptide p^k by:

$$u_k^h = \Pr(p^k \text{ presented}) = f\left(g_w(w^k;\theta_w) + g_h(x_h^i;\theta_h)\right), \quad (7)$$

where w^k denotes the encoded allele-noninteracting variables for peptide p^k, $g_w(\cdot)$ is a function for the allele-noninteracting variables w^k based on a set of parameters θ_w determined for the allele-noninteracting variables. Specifically, the values for the set of parameters θ_h for each MHC allele h and the set of parameters θ_w for allele-noninteracting variables can be determined by minimizing the loss function with respect to θ_h and θ_w, where i is each instance in the subset S of training data 170 generated from cells expressing single MHC alleles.

[00348] The output of the dependency function $g_w(w^k;\theta_w)$ represents a dependency score for the allele noninteracting variables indicating whether the peptide p^k will be presented by one or more MHC alleles based on the impact of allele noninteracting variables. For example, the dependency score for the allele noninteracting variables may have a high value if the peptide p^k is associated with a C-terminal flanking sequence that is known to positively impact presentation of the peptide p^k, and may have a low value if the peptide p^k is associated with a C-terminal flanking sequence that is known to negatively impact presentation of the peptide p^k.

[00349] According to equation (7), the per-allele likelihood that a peptide sequence p^k will be presented by a MHC allele h can be generated by applying the function $g_h(\cdot)$ for the MHC allele h to the encoded version of the peptide sequence p^k to generate the corresponding dependency score for allele interacting variables. The function $g_w(\cdot)$ for the allele noninteracting variables are also applied to the encoded version of the allele noninteracting variables to generate the dependency score for the allele noninteracting variables. Both scores are combined, and the combined score is transformed by the transformation function

$f(\cdot)$ to generate a per-allele likelihood that the peptide sequence p^k will be presented by the MHC allele h.

[00350] Alternatively, the training module 316 may include allele-noninteracting variables w^k in the prediction by adding the allele-noninteracting variables w^k to the allele-interacting variables x_h^k in equation (2). Thus, the presentation likelihood can be given by:

$$u_k^h = \Pr(\boldsymbol{p}^k \text{ presented; allele } h) = f\left(g_h([\boldsymbol{x}_h^k\ \boldsymbol{w}^k]; \boldsymbol{\theta}_h)\right). \quad (8)$$

VIII.B.3 Dependency Functions for Allele-Noninteracting Variables

[00351] Similarly to the dependency function $g_h(\cdot)$ for allele-interacting variables, the dependency function $g_w(\cdot)$ for allele noninteracting variables may be an affine function or a network function in which a separate network model is associated with allele-noninteracting variables w^k.

[00352] Specifically, the dependency function $g_w(\cdot)$is an affine function given by:

$$g_w(\boldsymbol{w}^k; \boldsymbol{\theta}_w) = \boldsymbol{w}^k \cdot \boldsymbol{\theta}_w.$$

that linearly combines the allele-noninteracting variables in w^k with a corresponding parameter in the set of parameters $\boldsymbol{\theta}_w$.

[00353] The dependency function $g_w(\cdot)$ may also be a network function given by:

$$g_w(\boldsymbol{w}^k; \boldsymbol{\theta}_w) = NN_w(\boldsymbol{w}^k; \boldsymbol{\theta}_w).$$

represented by a network model $NN_w(\cdot)$ having an associated parameter in the set of parameters $\boldsymbol{\theta}_w$. The network function may also include one or more network models each taking different allele noninteracting variables as input.

[00354] In another instance, the dependency function $g_w(\cdot)$ for the allele-noninteracting variables can be given by:

$$g_w(\boldsymbol{w}^k; \boldsymbol{\theta}_w) = g'_w(\boldsymbol{w}^k; \boldsymbol{\theta}'_w) + h(m^k; \theta_w^m), \quad (9)$$

where $g'_w(w^k;\theta'_w)$ is the affine function, the network function with the set of allele noninteracting parameters $\boldsymbol{\theta}'_w$, or the like, m^k is the mRNA quantification measurement for peptide p^k, $h(\cdot)$ is a function transforming the quantification measurement, and θ_w^m is a parameter in the set of parameters for allele noninteracting variables that is combined with the mRNA quantification measurement to generate a dependency score for the mRNA quantification measurement. In one particular embodiment referred throughout the remainder of the specification, $h(\cdot)$ is the log function, however in practice $h(\cdot)$ may be any one of a variety of different functions.

[00355] In yet another instance, the dependency function $g_w(\cdot)$ for the allele-noninteracting variables can be given by:

$$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w) = g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w) + \boldsymbol{\theta}_w^o \cdot \boldsymbol{o}^k, \tag{10}$$

where $g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)$ is the affine function, the network function with the set of allele noninteracting parameters $\boldsymbol{\theta}'_w$, or the like, $\boldsymbol{o}^k$ is the indicator vector described in Section VII.C.2 representing proteins and isoforms in the human proteome for peptide $\boldsymbol{p}^k$, and $\boldsymbol{\theta}_w^o$ is a set of parameters in the set of parameters for allele noninteracting variables that is combined with the indicator vector. In one variation, when the dimensionality of $\boldsymbol{o}^k$ and the set of parameters $\boldsymbol{\theta}_w^o$ are significantly high, a parameter regularization term, such as $\lambda \cdot ||\boldsymbol{\theta}_w^o||$, where $||\cdot||$ represents L1 norm, L2 norm, a combination, or the like, can be added to the loss function when determining the value of the parameters. The optimal value of the hyperparameter λ can be determined through appropriate methods.

[00356] In yet another instance, the dependency function $g_w(\cdot)$ for the allele-noninteracting variables can be given by:

$$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w) = g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w) + \sum_{l=1}^{L} \mathbb{1}(\text{gene}(\boldsymbol{p}^k) = l) \cdot \boldsymbol{\theta}_w^l, \tag{11}$$

where $g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)$ is the affine function, the network function with the set of allele noninteracting parameters $\boldsymbol{\theta}'_w$, or the like, $\mathbb{1}(\text{gene}(\boldsymbol{p}^k=l))$ is the indicator function that equals to 1 if peptide $\boldsymbol{p}^k$ is from source gene l as described above in reference to allele noninteracting variables, and $\boldsymbol{\theta}_w^l$ is a parameter indicating "antigenicity" of source gene l. In one variation, when L is significantly high, and thus, the number of parameters $\boldsymbol{\theta}_w^{l=1, 2, \ldots, L}$ are significantly high, a parameter regularization term, such as $\lambda \cdot ||\boldsymbol{\theta}_w^l||$, where $||\cdot||$ represents L1 norm, L2 norm, a combination, or the like, can be added to the loss function when determining the value of the parameters. The optimal value of the hyperparameter λ can be determined through appropriate methods.

[00357] In yet another instance, the dependency function $g_w(\cdot)$ for the allele-noninteracting variables can be given by:

$$g_w(\boldsymbol{w}^k;\boldsymbol{\theta}_w) = g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w) + \sum_{m=1}^{M}\sum_{l=1}^{L} \mathbb{1}(\text{gene}(\boldsymbol{p}^k) = l, \text{tissue}(\boldsymbol{p}^k) = m) \cdot \boldsymbol{\theta}_w^{lm}, \tag{12a}$$

where $g'_w(\boldsymbol{w}^k;\boldsymbol{\theta}'_w)$ is the affine function, the network function with the set of allele noninteracting parameters $\boldsymbol{\theta}'_w$, or the like, $\mathbb{1}(\text{gene}(\boldsymbol{p}^k)=l, \text{tissue}(\boldsymbol{p}^k)=m)$ is the indicator function that equals to 1 if peptide $\boldsymbol{p}^k$ is from source gene l and if peptide $\boldsymbol{p}^k$ is from tissue

type m as described above in reference to allele noninteracting variables, and θ_w^{lm} is a parameter indicating antigenicity of the combination of source gene l and tissue type m. Specifically, the antigenicity of gene l for tissue type m may denote the residual propensity for cells of tissue type m to present peptides from gene l after controlling for RNA expression and peptide sequence context.

[00358] In one variation, when L or M is significantly high, and thus, the number of parameters $\theta_w^{lm=1, 2, \ldots, LM}$ are significantly high, a parameter regularization term, such as as $\lambda \cdot ||\theta_w^{lm}||$, where $||\cdot||$ represents L1 norm, L2 norm, a combination, or the like, can be added to the loss function when determining the value of the parameters. The optimal value of the hyperparameter λ can be determined through appropriate methods. In another variation, a parameter regularization term can be added to the loss function when determining the value of the parameters, such that the parameters for the same source gene do not significantly differ between tissue types. For example, a penalization term such as:

$$\lambda \cdot \sum_{l=1}^{L} \sqrt{\sum_{m=1}^{M} \left(\theta_w^{lm} - \overline{\theta_w^{l}}\right)^2}$$

where $\overline{\theta_w^{l}}$ is the average antigenicity across tissue types for source gene l, may penalize the standard deviation of antigenicity across different tissue types in the loss function.

[00359] In yet another instance, the dependency function $g_w(\cdot)$ for the allele-noninteracting variables can be given by:

$$g_w(\boldsymbol{w}^k; \theta_w) = g'_w(\boldsymbol{w}^k; \theta'_w) + \sum_{l=1}^{L} \mathbb{1}(\text{gene}(p^k) = l) \cdot \theta_w^l + \sum_{m=1}^{M} \mathbb{1}(\text{loc}(p^k) = m) \cdot \theta_w^m \qquad (12b)$$

where $g'_w(\boldsymbol{w}^k; \theta'_w)$ is the affine function, the network function with the set of allele noninteracting parameters θ'_w, or the like, $\mathbb{1}(\text{gene}(p^k=l))$ is the indicator function that equals to 1 if peptide p^k is from source gene l as described above in reference to allele noninteracting variables, and θ_w^l is a parameter indicating "antigenicity" of source gene l, and $\mathbb{1}(\text{loc}(p^k=m))$ is the indicator function that equals to 1 if peptide p^k is from proteomic location m, and θ_w^m is a parameter indicating the extent to which proteomic location m is a presentation "hotspot". In one embodiment, a proteomic location can comprise a block of n

adjacent peptides from the same protein, where n is a hyperparameter of the model determined via appropriate methods such as grid-search cross-validation.

[00360] In practice, the additional terms of any of equations (9), (10), (11), (12a) and (12b) may be combined to generate the dependency function $g_w(\cdot)$ for allele noninteracting variables. For example, the term $h(\cdot)$ indicating mRNA quantification measurement in equation (9) and the term indicating source gene antigenicity in equations (11), (12a), and (12b) may be summed together along with any other affine or network function to generate the dependency function for allele noninteracting variables.

[00361] Returning to equation (7), as an example, the likelihood that peptide p^k will be presented by MHC allele $h=3$, among $m=4$ different identified MHC alleles using the affine transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k^3 = f\left(\boldsymbol{w}^k \cdot \boldsymbol{\theta}_w + \boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3\right),$$

where $\boldsymbol{w}^k$ are the identified allele-noninteracting variables for peptide p^k, and $\boldsymbol{\theta}_w$ are the set of parameters determined for the allele-noninteracting variables.

[00362] As another example, the likelihood that peptide p^k will be presented by MHC allele $h=3$, among $m=4$ different identified MHC alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k^3 = f\left(NN_w\left(\boldsymbol{w}^k; \boldsymbol{\theta}_w\right) + NN_3\left(\boldsymbol{x}_3^k; \boldsymbol{\theta}_3\right)\right)$$

where $\boldsymbol{w}^k$ are the identified allele-interacting variables for peptide p^k, and $\boldsymbol{\theta}_w$ are the set of parameters determined for allele-noninteracting variables.

[00363] FIG. 8 illustrates generating a presentation likelihood for peptide p^k in association with MHC allele $h=3$ using example network models $NN_3(\cdot)$ and $NN_w(\cdot)$. As shown in FIG. 8, the network model $NN_3(\cdot)$ receives the allele-interacting variables $\boldsymbol{x}_3^k$ for MHC allele $h=3$ and generates the output $NN_3(\boldsymbol{x}_3^k)$. The network model $NN_w(\cdot)$ receives the allele-noninteracting variables $\boldsymbol{w}^k$ for peptide p^k and generates the output $NN_w(\boldsymbol{w}^k)$. The outputs are combined and mapped by function $f(\cdot)$ to generate the estimated presentation likelihood u_k.

VIII.C. Multiple-Allele Models

[00364] The training module 316 may also construct the presentation models to predict presentation likelihoods of peptides in a multiple-allele setting where two or more MHC alleles are present. In this case, the training module 316 may train the presentation models based on data instances S in the training data 170 generated from cells expressing single MHC alleles, cells expressing multiple MHC alleles, or a combination thereof.

VIII.C.1. Example 1: Maximum of Per-Allele Models

[00365] In one implementation, the training module 316 models the estimated presentation likelihood u_k for peptide p^k in association with a set of multiple MHC alleles H as a function of the presentation likelihoods $u_k^{h \in H}$ determined for each of the MHC alleles h in the set H determined based on cells expressing single-alleles, as described above in conjunction with equations (2)-(10). Specifically, the presentation likelihood u_k can be any function of $u_k^{h \in H}$. In one implementation, as shown in equations (11), (12a), and (12b), the function is the maximum function, and the presentation likelihood u_k can be determined as the maximum of the presentation likelihoods for each MHC allele h in the set H.

$$u_k = \Pr(\boldsymbol{p}^k \text{ presented; alleles } H) = \max(u_k^{h \in H}).$$

VIII.C.2. Example 2.1: Function-of-Sums Models

[00366] In one implementation, the training module 316 models the estimated presentation likelihood u_k for peptide p^k by:

$$u_k = \Pr(\boldsymbol{p}^k \text{ presented}) = f\left(\sum_{h=1}^{m} a_h^k \cdot g_h(\boldsymbol{x}_h^k; \boldsymbol{\theta}_h)\right), \tag{13}$$

where elements a_h^k are 1 for the multiple MHC alleles H associated with peptide sequence p^k and x_h^k denotes the encoded allele-interacting variables for peptide p^k and the corresponding MHC alleles. The values for the set of parameters $\boldsymbol{\theta}_h$ for each MHC allele h can be determined by minimizing the loss function with respect to $\boldsymbol{\theta}_h$, where i is each instance in the subset S of training data 170 generated from cells expressing single MHC alleles and/or cells expressing multiple MHC alleles. The dependency function g_h may be in the form of any of the dependency functions g_h introduced above in sections VIII.B.1.

[00367] According to equation (13), the presentation likelihood that a peptide sequence p^k will be presented by one or more MHC alleles h can be generated by applying the dependency function $g_h(\cdot)$ to the encoded version of the peptide sequence p^k for each of the MHC alleles H to generate the corresponding score for the allele interacting variables. The scores for each MHC allele h are combined, and transformed by the transformation function $f(\cdot)$ to generate the presentation likelihood that peptide sequence p^k will be presented by the set of MHC alleles H.

[00368] The presentation model of equation (13) is different from the per-allele model of equation (2), in that the number of associated alleles for each peptide p^k can be greater than 1.

In other words, more than one element in a_h^k can have values of 1 for the multiple MHC alleles H associated with peptide sequence p^k.

[00369] As an example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the affine transformation functions $g_h(\cdot)$, can be generated by:

$$u_k = f\left(x_2^k \cdot \theta_2 + x_3^k \cdot \theta_3\right),$$

where x_2^k, x_3^k are the identified allele-interacting variables for MHC alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for MHC alleles $h=2$, $h=3$.

[00370] As another example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = f\left(NN_2(x_2^k; \theta_2) + NN_3(x_3^k; \theta_3)\right),$$

where $NN_2(\cdot)$, $NN_3(\cdot)$ are the identified network models for MHC alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for MHC alleles $h=2$, $h=3$.

[00371] FIG. 9 illustrates generating a presentation likelihood for peptide p^k in association with MHC alleles $h=2$, $h=3$ using example network models $NN_2(\cdot)$ and $NN_3(\cdot)$. As shown in FIG. 9, the network model $NN_2(\cdot)$ receives the allele-interacting variables x_2^k for MHC allele $h=2$ and generates the output $NN_2(x_2^k)$ and the network model $NN_3(\cdot)$ receives the allele-interacting variables x_3^k for MHC allele $h=3$ and generates the output $NN_3(x_3^k)$. The outputs are combined and mapped by function $f(\cdot)$ to generate the estimated presentation likelihood u_k.

VIII.C.3. Example 2.2: Function-of-Sums Models with Allele-Noninteracting Variables

[00372] In one implementation, the training module 316 incorporates allele-noninteracting variables and models the estimated presentation likelihood u_k for peptide p^k by:

$$u_k = \Pr\left(p^k \text{ presented}\right) = f\left(g_w(w^k; \theta_w) + \sum_{h=1}^{m} a_h^k \cdot g_h\left(x_h^k; \theta_h\right)\right), \quad (14)$$

where w^k denotes the encoded allele-noninteracting variables for peptide p^k. Specifically, the values for the set of parameters θ_h for each MHC allele h and the set of parameters θ_w for allele-noninteracting variables can be determined by minimizing the loss function with respect to θ_h and θ_w, where i is each instance in the subset S of training data 170 generated from cells expressing single MHC alleles and/or cells expressing multiple MHC alleles. The

dependency function g_w may be in the form of any of the dependency functions g_w introduced above in sections VIII.B.3.

[00373] Thus, according to equation (14), the presentation likelihood that a peptide sequence p^k will be presented by one or more MHC alleles H can be generated by applying the function $g_h(\cdot)$ to the encoded version of the peptide sequence p^k for each of the MHC alleles H to generate the corresponding dependency score for allele interacting variables for each MHC allele h. The function $g_w(\cdot)$ for the allele noninteracting variables is also applied to the encoded version of the allele noninteracting variables to generate the dependency score for the allele noninteracting variables. The scores are combined, and the combined score is transformed by the transformation function $f(\cdot)$ to generate the presentation likelihood that peptide sequence p^k will be presented by the MHC alleles H.

[00374] In the presentation model of equation (14), the number of associated alleles for each peptide p^k can be greater than 1. In other words, more than one element in a_h^k can have values of 1 for the multiple MHC alleles H associated with peptide sequence p^k.

[00375] As an example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the affine transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = f\left(\boldsymbol{w}^k \cdot \boldsymbol{\theta}_w + \boldsymbol{x}_2^k \cdot \boldsymbol{\theta}_2 + \boldsymbol{x}_3^k \cdot \boldsymbol{\theta}_3\right),$$

where $\boldsymbol{w}^k$ are the identified allele-noninteracting variables for peptide p^k, and $\boldsymbol{\theta}_w$ are the set of parameters determined for the allele-noninteracting variables.

[00376] As another example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = f\left(NN_w\left(\boldsymbol{w}^k; \boldsymbol{\theta}_w\right) + NN_2(\boldsymbol{x}_2^k; \boldsymbol{\theta}_2) + NN_3(\boldsymbol{x}_3^k; \boldsymbol{\theta}_3)\right)$$

where $\boldsymbol{w}^k$ are the identified allele-interacting variables for peptide p^k, and $\boldsymbol{\theta}_w$ are the set of parameters determined for allele-noninteracting variables.

[00377] FIG. 10 illustrates generating a presentation likelihood for peptide p^k in association with MHC alleles $h=2$, $h=3$ using example network models $NN_2(\cdot)$, $NN_3(\cdot)$, and $NN_w(\cdot)$. As shown in FIG. 10, the network model $NN_2(\cdot)$ receives the allele-interacting variables $\boldsymbol{x}_2^k$ for MHC allele $h=2$ and generates the output $NN_2(\boldsymbol{x}_2^k)$. The network model $NN_3(\cdot)$ receives the allele-interacting variables $\boldsymbol{x}_3^k$ for MHC allele $h=3$ and generates the output $NN_3(\boldsymbol{x}_3^k)$. The network model $NN_w(\cdot)$ receives the allele-noninteracting variables $\boldsymbol{w}^k$

for peptide p^k and generates the output $NN_w(w^k)$. The outputs are combined and mapped by function $f(\cdot)$ to generate the estimated presentation likelihood u_k.

[00378] Alternatively, the training module 316 may include allele-noninteracting variables w^k in the prediction by adding the allele-noninteracting variables w^k to the allele-interacting variables x_h^k in equation (15). Thus, the presentation likelihood can be given by:

$$u_k = \Pr(p^k \text{ presented}) = f\left(\sum_{h=1}^{m} a_h^k \cdot g_h([x_h^k\, w^k]; \theta_h)\right). \quad (15)$$

VIII.C.4. Example 3.1: Models Using Implicit Per-Allele Likelihoods

[00379] In another implementation, the training module 316 models the estimated presentation likelihood u_k for peptide p^k by:

$$u_k = \Pr(p^k \text{ presented}) = r\left(s\left(v = \left[a_1^k \cdot u'^1_k(\theta) \ldots a_m^k \cdot u'^m_k(\theta)\right]\right)\right), \quad (16)$$

where elements a_h^k are 1 for the multiple MHC alleles $h \in H$ associated with peptide sequence p^k, u'^h_k is an implicit per-allele presentation likelihood for MHC allele h, vector v is a vector in which element v_h corresponds to $a_h^k \cdot u'^h_k$, $s(\cdot)$ is a function mapping the elements of v, and $r(\cdot)$ is a clipping function that clips the value of the input into a given range. As described below in more detail, $s(\cdot)$ may be the summation function or the second-order function, but it is appreciated that in other embodiments, $s(\cdot)$ can be any function such as the maximum function. The values for the set of parameters θ for the implicit per-allele likelihoods can be determined by minimizing the loss function with respect to θ, where i is each instance in the subset S of training data 170 generated from cells expressing single MHC alleles and/or cells expressing multiple MHC alleles.

[00380] The presentation likelihood in the presentation model of equation (16) is modeled as a function of implicit per-allele presentation likelihoods u'^h_k that each correspond to the likelihood peptide p^k will be presented by an individual MHC allele h. The implicit per-allele likelihood is distinct from the per-allele presentation likelihood of section VIII.B in that the parameters for implicit per-allele likelihoods can be learned from multiple allele settings, in which direct association between a presented peptide and the corresponding MHC allele is unknown, in addition to single-allele settings. Thus, in a multiple-allele setting, the presentation model can estimate not only whether peptide p^k will be presented by a set of MHC alleles H as a whole, but can also provide individual likelihoods $u'^{h \in H}_k$ that indicate which MHC allele h most likely presented peptide p^k. An advantage of this is that the

presentation model can generate the implicit likelihoods without training data for cells expressing single MHC alleles.

[00381] In one particular implementation referred throughout the remainder of the specification, $r(\cdot)$ is a function having the range [0, 1]. For example, $r(\cdot)$ may be the clip function:

$$r(z) = \min(\max(z, 0), 1),$$

where the minimum value between z and 1 is chosen as the presentation likelihood u_k. In another implementation, $r(\cdot)$ is the hyperbolic tangent function given by:

$$r(z) = \tanh(z)$$

when the values for the domain z is equal to or greater than 0.

VIII.C.5. Example 3.2: Sum-of-Functions Model

[00382] In one particular implementation, $s(\cdot)$ is a summation function, and the presentation likelihood is given by summing the implicit per-allele presentation likelihoods:

$$u_k = \Pr(\boldsymbol{p}^k \text{ presented}) = r\left(\sum_{h=1}^{m} a_h^k \cdot u'^h_k(\boldsymbol{\theta})\right). \quad (17)$$

[00383] In one implementation, the implicit per-allele presentation likelihood for MHC allele h is generated by:

$$u'^h_k = f\left(g_h(\boldsymbol{x}_h^k; \boldsymbol{\theta}_h)\right), \quad (18)$$

such that the presentation likelihood is estimated by:

$$u_k = \Pr(\boldsymbol{p}^k \text{ presented}) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_h(\boldsymbol{x}_h^k; \boldsymbol{\theta}_h)\right)\right). \quad (19)$$

[00384] According to equation (19), the presentation likelihood that a peptide sequence $\boldsymbol{p}^k$ will be presented by one or more MHC alleles H can be generated by applying the function $g_h(\cdot)$ to the encoded version of the peptide sequence $\boldsymbol{p}^k$ for each of the MHC alleles H to generate the corresponding dependency score for allele interacting variables. Each dependency score is first transformed by the function $f(\cdot)$ to generate implicit per-allele presentation likelihoods u'^h_k. The per-allele likelihoods u'^h_k are combined, and the clipping function may be applied to the combined likelihoods to clip the values into a range [0, 1] to generate the presentation likelihood that peptide sequence $\boldsymbol{p}^k$ will be presented by the set of MHC alleles H. The dependency function g_h may be in the form of any of the dependency functions g_h introduced above in sections VIII.B.1.

[00385] As an example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the affine transformation functions $g_h(\cdot)$, can be generated by:

$$u_k = r\left(f\left(x_2^k \cdot \theta_2\right) + f\left(x_3^k \cdot \theta_3\right)\right),$$

where x_2^k, x_3^k are the identified allele-interacting variables for MHC alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for MHC alleles $h=2$, $h=3$.

[00386] As another example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = r\left(f\left(NN_2(x_2^k; \theta_2)\right) + f\left(NN_3(x_3^k; \theta_3)\right)\right),$$

where $NN_2(\cdot)$, $NN_3(\cdot)$ are the identified network models for MHC alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for MHC alleles $h=2$, $h=3$.

[00387] FIG. 11 illustrates generating a presentation likelihood for peptide p^k in association with MHC alleles $h=2$, $h=3$ using example network models $NN_2(\cdot)$ and $NN_3(\cdot)$. As shown in FIG. 11, the network model $NN_2(\cdot)$ receives the allele-interacting variables x_2^k for MHC allele $h=2$ and generates the output $NN_2(x_2^k)$ and the network model $NN_3(\cdot)$ receives the allele-interacting variables x_3^k for MHC allele $h=3$ and generates the output $NN_3(x_3^k)$. Each output is mapped by function $f(\cdot)$ and combined to generate the estimated presentation likelihood u_k.

[00388] In another implementation, when the predictions are made for the log of mass spectrometry ion currents, $r(\cdot)$ is the log function and $f(\cdot)$ is the exponential function.

VIII.C.6. Example 3.3: Sum-of-Functions Models with Allele-noninteracting Variables

[00389] In one implementation, the implicit per-allele presentation likelihood for MHC allele h is generated by:

$$u_k'^{h} = f\left(g_h(x_h^k; \theta_h) + g_w(w^k; \theta_w)\right), \tag{20}$$

such that the presentation likelihood is generated by:

$$u_k = \Pr(p^k \text{ presented}) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_w(w^k; \theta_w) + g_h(x_h^k; \theta_h)\right)\right), \tag{21}$$

to incorporate the impact of allele noninteracting variables on peptide presentation.

[00390] According to equation (21), the presentation likelihood that a peptide sequence p^k will be presented by one or more MHC alleles H can be generated by applying the function

$g_h(\cdot)$ to the encoded version of the peptide sequence p^k for each of the MHC alleles H to generate the corresponding dependency score for allele interacting variables for each MHC allele h. The function $g_w(\cdot)$ for the allele noninteracting variables is also applied to the encoded version of the allele noninteracting variables to generate the dependency score for the allele noninteracting variables. The score for the allele noninteracting variables are combined to each of the dependency scores for the allele interacting variables. Each of the combined scores are transformed by the function $f(\cdot)$ to generate the implicit per-allele presentation likelihoods. The implicit likelihoods are combined, and the clipping function may be applied to the combined outputs to clip the values into a range [0,1] to generate the presentation likelihood that peptide sequence p^k will be presented by the MHC alleles H. The dependency function g_w may be in the form of any of the dependency functions g_w introduced above in sections VIII.B.3.

[00391] As an example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the affine transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = r\left(f\left(w^k \cdot \theta_w + x_2^k \cdot \theta_2\right) + f\left(w^k \cdot \theta_w + x_3^k \cdot \theta_3\right)\right),$$

where w^k are the identified allele-noninteracting variables for peptide p^k, and θ_w are the set of parameters determined for the allele-noninteracting variables.

[00392] As another example, the likelihood that peptide p^k will be presented by MHC alleles $h=2$, $h=3$, among $m=4$ different identified MHC alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = r\left(f\left(NN_w\left(w^k;\theta_w\right) + NN_2(x_2^k;\theta_2)\right) + f\left(NN_w\left(w^k;\theta_w\right) + NN_3(x_3^k;\theta_3)\right)\right)$$

where w^k are the identified allele-interacting variables for peptide p^k, and θ_w are the set of parameters determined for allele-noninteracting variables.

[00393] FIG. 12 illustrates generating a presentation likelihood for peptide p^k in association with MHC alleles $h=2$, $h=3$ using example network models $NN_2(\cdot)$, $NN_3(\cdot)$, and $NN_w(\cdot)$. As shown in FIG. 12, the network model $NN_2(\cdot)$ receives the allele-interacting variables x_2^k for MHC allele $h=2$ and generates the output $NN_2(x_2^k)$. The network model $NN_w(\cdot)$ receives the allele-noninteracting variables w^k for peptide p^k and generates the output $NN_w(w^k)$. The outputs are combined and mapped by function $f(\cdot)$. The network model $NN_3(\cdot)$ receives the allele-interacting variables x_3^k for MHC allele $h=3$ and generates the output $NN_3(x_3^k)$, which is again combined with the output $NN_w(w^k)$ of the same network model

$NN_w(\cdot)$ and mapped by function $f(\cdot)$. Both outputs are combined to generate the estimated presentation likelihood u_k.

[00394] In another implementation, the implicit per-allele presentation likelihood for MHC allele h is generated by:

$$u_k'^h = f\left(g_h\left([x_h^k\, w^k]; \theta_h\right)\right). \tag{22}$$

such that the presentation likelihood is generated by:

$$u_k = \Pr(p^k \text{ presented}) = r\left(\sum_{h=1}^{m} a_h^k \cdot f\left(g_h\left([x_h^k\, w^k]; \theta_h\right)\right)\right).$$

VIII.C.7. Example 4: Second Order Models

[00395] In one implementation, , $s(\cdot)$ is a second-order function, and the estimated presentation likelihood u_k for peptide p^k is given by:

$$u_k = \Pr(p^k \text{ presented}) = \sum_{h=1}^{m} a_h^k \cdot u_k'^h(\theta) - \sum_{h=1}^{m}\sum_{j<h} a_h^k \cdot a_j^k \cdot u_k'^h(\theta) \cdot u_k'^j(\theta) \tag{23}$$

where elements u'^h_k are the implicit per-allele presentation likelihood for MHC allele h. The values for the set of parameters θ for the implicit per-allele likelihoods can be determined by minimizing the loss function with respect to θ, where i is each instance in the subset S of training data 170 generated from cells expressing single MHC alleles and/or cells expressing multiple MHC alleles. The implicit per-allele presentation likelihoods may be in any form shown in equations (18), (20), and (22) described above.

[00396] In one aspect, the model of equation (23) may imply that there exists a possibility peptide p^k will be presented by two MHC alleles simultaneously, in which the presentation by two HLA alleles is statistically independent.

[00397] According to equation (23), the presentation likelihood that a peptide sequence p^k will be presented by one or more MHC alleles H can be generated by combining the implicit per-allele presentation likelihoods and subtracting the likelihood that each pair of MHC alleles will simultaneously present the peptide p^k from the summation to generate the presentation likelihood that peptide sequence p^k will be presented by the MHC alleles H.

[00398] As an example, the likelihood that peptide p^k will be presented by HLA alleles $h=2$, $h=3$, among $m=4$ different identified HLA alleles using the affine transformation functions $g_h(\cdot)$, can be generated by:

$$u_k = f(x_2^k \cdot \theta_2) + f(x_3^k \cdot \theta_3) - f(x_2^k \cdot \theta_2) \cdot f(x_3^k \cdot \theta_3),$$

where x_2^k, x_3^k are the identified allele-interacting variables for HLA alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for HLA alleles $h=2$, $h=3$.

[00399] As another example, the likelihood that peptide p^k will be presented by HLA alleles $h=2$, $h=3$, among $m=4$ different identified HLA alleles using the network transformation functions $g_h(\cdot)$, $g_w(\cdot)$, can be generated by:

$$u_k = f\left(NN_2(x_2^k;\theta_2)\right) + f\left(NN_3(x_3^k;\theta_3)\right) - f\left(NN_2(x_2^k;\theta_2)\right) \cdot f\left(NN_3(x_3^k;\theta_3)\right),$$

where $NN_2(\cdot)$, $NN_3(\cdot)$ are the identified network models for HLA alleles $h=2$, $h=3$, and θ_2, θ_3 are the set of parameters determined for HLA alleles $h=2$, $h=3$.

VIII.D. Pan-Allele Models

[00400] In contrast to the per-allele model, a pan-allele model is a presentation model that is capable of predicting presentation likelihoods of peptides on a pan-allele basis. Specifically, unlike the per-allele model that is capable of predicting the probability that peptides will be presented by one or more known MHC alleles that have been previously used to train the per-allele model, the pan-allele model is a presentation model that is capable of predicting the probability that a peptide will be presented by *any* MHC allele—including unknown MHC alleles that the model has not previously encountered during training.

[00401] Briefly, the pan-allele model is trained by the training module 316. Similar to the training of the per-allele model, the training module 316 may train the pan-allele presentation model based on data instances S in the training data 170 generated from cells expressing single MHC alleles, cells expressing multiple MHC alleles, or a combination thereof. However rather than training the pan-allele presentation model using a particular MHC allele or a particular set of MHC alleles a^k_h, the training module 316 trains the pan-allele presentation model using all MHC allele peptide sequences d_h available in the training data 170. Specifically, the training module 316 trains the pan-allele presentation model based on positions of amino acids of the MHC alleles available in the training data 170.

[00402] After the pan-allele model has been trained, when a peptide sequence and known or unknown MHC allele peptide sequence are input into the model to determine the probability that the known or unknown MHC allele will present the peptide, the model is able to accurately predict this probability by using information learned during training with similar MHC allele peptide sequences. For example, a pan-allele model trained using training data 170 that does not contain any occurrences of the A*02:07 allele may still accurately predict the presentation of peptides by the A*02:07 allele by drawing upon information learned during training with similar alleles (e.g., alleles in the A*02 gene family). In this

way, a single presentation pan-allele model can predict presentation likelihoods of a peptide on any MHC allele.

VIII.D.2. Advantages of Pan-Allele Models

[00403] The principle advantage of the pan-allele presentation model is that the pan-allele presentation model has greater versatility than the per-allele presentation model. As noted above, a per-allele model is capable of predicting the probability that a peptide will be presented by one or more identified MHC alleles that were used to train the per-allele model. In other words, the per-allele model is associated with a limited set of one or more known MHC alleles.

[00404] Therefore, given a sample containing a particular set of one or more MHC alleles, to determine the probability that a peptide is presented by the particular set of MHC alleles, a per-allele model that was trained using that particular set of MHC alleles is selected for use. In other words, when relying on per-allele models to predict the probability that a peptide will be presented by an MHC allele, predictions can be made only for MHC alleles that have appeared in the training data 170. Because a large number of MHC alleles exist (particularly for minor variations within the same gene family), a very large quantity of training samples would be required to train per-allele presentation models to be equipped make peptide presentation predictions for all MHC alleles.

[00405] In contrast, the pan-allele model is not limited to making predictions for a particular set of one or more MHC alleles on which it was trained. Instead, during use, the pan-allele model is able to accurately predict the probability that a previously-seen and/or a previously-unseen MHC allele will present a given peptide by using information learned during training with similar MHC allele peptide sequences. As a result, the pan-allele model is not associated with a particular set of one or more MHC alleles, and is capable of predicting the probability that a peptide will be presented by *any* MHC allele. This versatility of the pan-allele model means that a single model can be used to predict the likelihood that any peptide will be presented by any MHC allele. Therefore, use of the pan-allele model reduces the amount of training data required to maximize both individual HLA coverage and population HLA coverage, as defined above in Section VII.A..

VIII.D.3. Use of Pan-Allele Models

[00406] The following discussion in sections VIII.D.4. – VIII.D.7. concerns use of the pan-allele model to predict the probability that a peptide will be presented by one or more

MHC allele(s). For simplicity, this discussion operates under the assumption that the pan-allele model has already been trained by the training module 316. Training of the pan-allele model is discussed in detail below with regard to section VIII.D.8..

[00407] Furthermore, the following discussion in Sections VIII.D.4. – VIII.D.6. pertains to use of the pan-allele model to predict the likelihood that a peptide will be presented by a single MHC allele and/or by multiple MHC alleles in a given sample. However, as described in further detail below with regard to Section VIII.D.7. there are slight differences between using the pan-allele model predict the likelihood that a peptide will be presented by a single MHC allele in a sample and using the pan-allele model to predict the likelihood that a peptide will be presented by multiple MHC alleles in a sample.

[00408] Briefly, when using the pan-allele model to predict the likelihood that a peptide will be presented by a single MHC allele, one set of inputs is provided to the pan-allele model as described in detail below, and the pan-allele model generates a single output.

[00409] On the other hand, when using the pan-allele model to predict the likelihoods that a peptide will be presented by multiple MHC alleles, the pan-allele model is used iteratively for each MHC allele of the multiple MHC alleles. Specifically, when using the pan-allele model to predict the likelihoods that a peptide will be presented by multiple MHC alleles, a first set of inputs associated with a first MHC allele of the multiple MHC alleles is provided to the pan-allele model, and the pan-allele model generates a first output for the first MHC allele. Then, a second set of inputs associated with a second MHC allele of the multiple MHC alleles is provided to the pan-allele model, and the pan-allele model generates a second output for the second MHC allele. This process is performed iteratively for each MHC allele of the multiple MHC alleles. Finally, the outputs generated by the pan-allele model for each MHC allele of the multiple MHC alleles are combined to generate a single probability that the multiple MHC alleles present the given peptide as described with regard to Section VIII.D.7..

VIII.D.4. Overview of Pan-Allele Models

[00410] In one implementation, a pan-allele model is used to estimate the presentation likelihood u_k for peptide $\boldsymbol{p}^k$ for a allele h. In some embodiments, the pan-allele model is represented by the equation:

$$u_k^h = \Pr(\boldsymbol{p}^k \text{ presented; MHC allele } h) = f\left(g_H([\boldsymbol{p}^k\ \boldsymbol{d}_h]; \boldsymbol{\theta}_H)\right), \qquad (24)$$

where p^k denotes the peptide sequence, d_h denotes the peptide sequence of MHC allele h, $f(\cdot)$ is any transformation function, and $g_H(\cdot)$ is any dependency function. The pan-allele model generates dependency scores for the peptide sequence p^k and the MHC allele peptide sequence d_h based on a set of shared parameters θ_H determined for all MHC alleles. The values of the set of shared parameters θ_H are learned during training of the pan-allele model and are discussed in detail below in section VIII.D.8..

[00411] The output of the dependency function $g_H([p^k\ d_h];\theta_H)$ represents a dependency score for the MHC allele h indicating whether the MHC allele h will present the peptide p^k based on at least the positions of amino acids of the peptide sequence p^k and the positions of amino acids of the MHC allele peptide sequence d_h. For example, the dependency score for the MHC allele h may have a high value if the MHC allele h is likely to present the peptide p^k given an input MHC allele peptide sequence d_h, and may have a low value if presentation is not likely. The transformation function $f(\cdot)$ transforms the input, and more specifically, transforms the dependency score generated by $g_H([p^k\ d_h];\theta_H)$ in this case, to an appropriate value to indicate the likelihood that the peptide p^k will be presented by the MHC allele h.

[00412] In one particular implementation referred to throughout the remainder of the specification, $f(\cdot)$ is a function having the range within [0, 1] for an appropriate domain range. In one example, $f(\cdot)$ is the expit function. As another example, $f(\cdot)$ can also be the hyperbolic tangent function when the values for the domain z is equal to or greater than 0. Alternatively, when predictions are made for the mass spectrometry ion current that have values outside the range [0, 1], $f(\cdot)$ can be any function such as the identity function, the exponential function, the log function, and the like.

[00413] Thus, the likelihood that a peptide sequence p^k will be presented by a MHC allele h can be generated by applying the dependency function $g_H(\cdot)$ to the encoded version of the peptide sequence p^k and to the encoded version of the MCH allele peptide sequence d_h to generate the corresponding dependency score. The dependency score may be transformed by the transformation function $f(\cdot)$ to generate a likelihood that the peptide sequence p^k will be presented by the MHC allele h.

VIII.D.5. Dependency Functions for Allele-Interacting Variables

[00414] In one particular implementation referred to throughout the specification, the dependency function $g_H(\cdot)$ is an affine function given by:

$$g_H([p^k\ d_h];\theta_H) = \sum_{i=1}^{n_{pep}} \sum_{j=1}^{n_{MHC}} \sum_{k=1}^{20} \sum_{l=1}^{20} 1[p_i^k = k]1[d_{hj} = l]\theta_{H,ijkl} + \alpha, \quad (25)$$

where α is an intercept, p_i^k denotes the residue at position i of peptide p^k, d_{hj} denotes the residue at position j of MHC allele h, $1[]$ denotes an indicator variable whose value is 1 if the condition inside the brackets is true and 0 otherwise, $p_i^k = k$ is true if the amino acid at position i of peptide p^k is amino acid k and false otherwise, $d_{hj} = l$ is true if the amino acid at position j of MHC allele h is amino acid l and false otherwise, n_{pep} denotes the length of peptides modeled, n_{MHC} denotes the number of MHC residues considered in the model, and $\theta_{H,ijkl}$ is a coefficient describing the contribution of having residue k at position i of the peptide and residue l at position j of the MHC allele to the likelihood of presentation. This is a linear model in the one hot-encoded peptide sequence and the one hot-encoded MHC allele sequence, with peptide-residue-by-MHC-residue interactions for all peptide residues and MHC allele residues.

[00415] In another particular implementation referred to throughout the specification, the dependency function $g_H(\cdot)$ is a network function given by:

$$g_H([p^k\ d_h];\theta_H) = NN_H([p^k\ d_h];\theta_H) \quad (26)$$

represented by a network model $NN_H(\cdot)$ having a series of nodes arranged in one or more layers. A node may be connected to other nodes through connections each having an associated parameter in the set of parameters θ_H. A value at one particular node may be represented as a sum of the values of nodes connected to the particular node weighted by the associated parameter mapped by an activation function associated with the particular node. In contrast to the affine function, network models are advantageous because the presentation model can incorporate non-linearity and process data having different lengths of amino acid sequences. Specifically, through non-linear modeling, network models can capture interaction between amino acids at different positions in a peptide sequence, as well as interaction between amino acids at different positions in a MHC allele peptide sequence, and how these interactions affects peptide presentation.

[00416] In general, network models $NN_H(\cdot)$ may be structured as feed-forward networks, such as artificial neural networks (ANN), convolutional neural networks (CNN), deep neural networks (DNN), and/or recurrent networks, such as long short-term memory networks (LSTM), bi-directional recurrent networks, deep bi-directional recurrent networks, and the like.

[00417] In one instance, the single network model $NN_H(\cdot)$ may be a network model that outputs a dependency score given an encoded peptide sequence $\boldsymbol{p}^k$ and an encoded protein sequence $\boldsymbol{d}_h$ of an MHC allele h. In such an instance, the set of parameters $\boldsymbol{\theta}_H$ may correspond to a set of parameters for the single network model, and thus, the set of parameters $\boldsymbol{\theta}_H$ may be shared by all MHC alleles. Thus, in such an instance, $NN_H(\cdot)$ may denote the output of the single network model $NN_H(\cdot)$ given any inputs $[\boldsymbol{p}^k \boldsymbol{d}_h]$ to the single network model. As discussed above, such a network model is advantageous because peptide presentation probabilities for MHC alleles that were unknown in the training data can be predicted just by identification of the MHC alleles' protein sequences.

[00418] FIG. 13 illustrates an example network model $NN_H(\cdot)$ shared by MHC alleles. As shown in FIG. 13, the network model $NN_H(\cdot)$ receives the peptide sequence $\boldsymbol{p}^k$ and protein sequence $\boldsymbol{d}_h$ of an MHC allele h as input, and outputs a dependency score $NN_H([\boldsymbol{p}^k \boldsymbol{d}_h])$ corresponding to the MHC allele h.

[00419] FIG. 14 illustrates an example network model $NN_H(\cdot)$. As shown in FIG. 14, the network model $NN_H(\cdot)$ includes four input nodes at layer $l=1$, five nodes at layer $l=2$, two nodes at layer $l=3$, and one output node at layer $l=4$. In alternative embodiments, the network model $NN_H(\cdot)$ may contain any number of layers, and each layer may contain any number of nodes. The network model $NN_H(\cdot)$ is associated with a set of thirteen nonzero parameters $\theta_H(1)$, $\theta_H(2)$, ..., $\theta_H(13)$. These parameters serve to transform the values that are propagated from node to node, through the network model.

[00420] As shown in FIG. 14, the four input nodes at layer $l=1$ of the network model $NN_H(\cdot)$ receive input values including encoded polypeptide sequence data and encoded MHC allele peptide sequence data. The encoded polypeptide sequence data contains the amino acid sequence for a peptide, and the encoded MHC allele peptide sequence data contains the amino acid sequence for an MHC allele that may (or may not) present the peptide. In certain embodiments, once input into the network model $NN_H(\cdot)$ via the input nodes at layer $l=1$, the encoded polypeptide sequence is concatenated to the front of the encoded MHC allele peptide sequence within a layer of the network model $NN_H(\cdot)$. These input values are then propagated through the network model $NN_H(\cdot)$ according to the values of the parameters. In some embodiments, the layers of the network model $NN_H(\cdot)$ include two fully-connected dense network layers. In further embodiments, the first layer of these two fully-connected dense network layers comprises between 64-128 nodes with a rectified linear unit activation function. In even further embodiments, the second layer of these two fully-connected dense

network layers comprises a single node with a linear output. In such embodiments, this single node may be the output node of the network model $NN_H(\cdot)$. Finally, the network model $NN_H(\cdot)$ outputs the value $NN_H([p^k\ d_h])$. This output represents a dependency score for the MHC allele h indicating whether the MHC allele h will present the peptide sequence p^k. The network function may also include one or more network models each taking different allele-interacting variables (e.g., peptide sequences) as input.

[00421] In yet another instance, the dependency function $g_H(\cdot)$ can be expressed as:

$$g_H([p^k\ d_h];\theta_H) = g'_H([p^k\ d_h];\theta'_H) + \theta_H^0$$

where $g'_H([p^k\ d_h];\theta'_H)$ is the affine function with a set of parameters θ'_H, the network function, or the like, with a bias parameter θ_H^0 in the set of shared parameters θ_H for allele-interacting variables that represents a baseline probability of presentation for any MHC allele.

[00422] In another implementation, the bias parameter θ_H^0 may be shared according to the gene family of the MHC allele h. That is, the bias parameter θ_H^0 for MHC allele h may be equal to $\theta_{gene(h)}^0$, where $gene(h)$ is the gene family of MHC allele h. For example, class I MHC alleles HLA-A*02:01, HLA-A*02:02, and HLA-A*02:03 may be assigned to the gene family of "HLA-A," and the bias parameter θ_H^0 for each of these MHC alleles may be shared. As another example, class II MHC alleles HLA-DRB1:10:01, HLA-DRB1:11:01, and HLA-DRB3:01:01 may be assigned to the gene family of "HLA-DRB," and the bias parameter θ_H^0 for each of these MHC alleles may be shared. As discussed above, gene family may be one of the allele-interacting variables associated with an MHC allele h.

[00423] Returning to equation (23), as an example, the likelihood that peptide p^k will be presented by MHC allele h, using the affine dependency function $g_H(\cdot)$, can be generated by:

$$u_k^h = f\left(\sum_{i=1}^{n_{pep}}\sum_{j=1}^{n_{MHC}}\sum_{k=1}^{20}\sum_{l=1}^{20} 1[p_i^k = k]1[d_{hj} = l]\theta_{H,ijkl} + \alpha\right),$$

where α is an intercept, p_i^k denotes the residue at position i of peptide p^k, d_{hj} denotes the residue at position j of MHC allele h, $1[]$ denotes an indicator variable whose value is 1 if the condition inside the brackets is true and 0 otherwise, $p_i^k = k$ is true if the amino acid at position i of peptide p^k is amino acid k and false otherwise, $d_{hj} = l$ is true if the amino acid at position j of MHC allele h is amino acid l and false otherwise, n_{pep} denotes the length of peptides modeled, n_{MHC} denotes the number of MHC residues considered in the model, and $\theta_{H,ijkl}$ is a coefficient describing the contribution of having residue k at position i of the peptide and residue l at position j of the MHC allele to the likelihood of presentation. This is

a linear model in the one hot-encoded peptide sequence and the one hot-encoded MHC allele sequence, with peptide-residue-by-MHC-residue interactions for all peptide residues and MHC allele residues.

[00424] As another example, the likelihood that peptide p^k will be presented by an MHC allele h, using the network transformation function $g_H(\cdot)$, can be generated by:

$$u_k^h = f\left(NN_H([p^k\ d_h\ ]; \theta_H)\right),$$

where p^k denotes the peptide sequence, d_h denotes the peptide sequence of MHC allele h, and θ_H is the set of parameters determined for the network model $NN_H(\cdot)$ that is associated with all MHC alleles.

[00425] FIG. 15 illustrates generating a presentation likelihood for a peptide p^k in association with MHC allele h using an example shared network model $NN_H(\cdot)$. As shown in FIG. 15, the shared network model $NN_H(\cdot)$ receives the peptide sequence p^k and the MHC allele peptide sequence d_h, and generates the output $NN_H([p^k\ d_h])$. The output is mapped by function $f(\cdot)$ to generate the estimated presentation likelihood u_k.

VIII.D.6. Allele-Noninteracting Variables

[00426] As discussed above, allele-noninteracting variables comprise information that influences presentation of peptides that are independent of the type of MHC allele. For example, allele-noninteracting variables may include protein sequences on the N-terminus and C-terminus of the peptide, the protein family of the presented peptide, the level of RNA expression of the source gene of the peptides, and any additional allele-noninteracting variables.

[00427] In one implementation, the training module 316 incorporates allele-noninteracting variables into the pan-allele presentation models in a similar manner as described with regard to the per-allele models and the multiple allele models. For example, in some embodiments, allele-noninteracting variables may be entered as inputs into a dependency function that is separate from the dependency function used for allele-interacting variables. In such embodiments, the outputs of the two separate dependency functions may be summed, and the resulting summation may be input into the transformation function to generate a presentation prediction. Such embodiments for incorporating allele-noninteracting variables into pan-allele models, as well as others, are discussed above in sections VIII.B.2., VIII.B.3., VIII.C.3., and VIII.C.6..

VIII.D.7. Multiple-Allele Samples

[00428] As described above, a test sample may contain multiple MHC alleles rather than a single MHC allele. In fact, a majority of samples taken from nature include more than one MHC allele. For example, each human genome contains six MHC class I loci. Therefore, a sample that contains a human genome can contain up to six different MHC class I alleles. Accordingly, samples that contain multiple MHC alleles, rather than a single MHC allele, are typical samples of real-life test cases.

[00429] In embodiments in which a test sample contains multiple MHC alleles, the pan-allele model described above in Sections VIII.D.4. – VIII.D.6. may be employed to determine the probability that a given peptide from the test sample is presented by the multiple MHC alleles. However, as described briefly above, when using the pan-allele model to predict the likelihoods that a peptide will be presented by multiple MHC alleles, the pan-allele model described above is used *iteratively* for each MHC allele of the multiple MHC alleles. In other words, for each MHC allele of the multiple MHC alleles, the MHC allele peptide sequence and the peptide sequence are independently input into the dependency function shared by all MHC alleles. Based on these inputs, an output corresponding to the MHC allele is generated by the dependency function. This process is performed iteratively for each MHC allele of the multiple MHC alleles. Accordingly, each MHC allele of the multiple MHC alleles is independently associated with an output of the dependency function. The outputs associated with each MHC allele of the multiple MHC alleles are then combined.

[00430] The outputs of the dependency function that are associated with each MHC allele of the multiple MHC alleles can be combined as described with regard to sections VIII.C. – VIII.C.7.. As described with regard to sections VIII.C. – VIII.C.7., the manner in which the multiple outputs of the dependency function are combined can vary. For example, in some embodiments, the outputs of the dependency function iterations may be summed, and the resulting summation may be input into a transformation function to generate a presentation prediction. An equation that captures such an embodiment can be written as:

$$u_k^{\mathrm{h}} = \Pr\left(\boldsymbol{p}^k \text{ presented; MHC allele } h\right) = f\left(\sum_{h=1}^{T} g_H\left([\boldsymbol{p}^k\, \boldsymbol{d}_h\,]; \boldsymbol{\theta}_H\right)\right), \tag{27}$$

where T is the total number of unique MHC alleles in a sample containing multiple alleles. In alternative embodiments, the each individual output of the dependency function iterations may be input into a transformation function, and the resulting outputs from the

transformation functions may be summed to generate a presentation prediction. An equation that captures this alternative embodiment can be written as:

$$u_k^h = \Pr(\boldsymbol{p}^k \text{ presented; MHC allele } h) = \left(\sum_{h=1}^{T} f\left(g_H([\boldsymbol{p}^k\, \boldsymbol{d}_h]; \boldsymbol{\theta}_H)\right)\right), \quad (28)$$

Such embodiments, as well as others, in which multiple outputs of the dependency function are combined to predict the probability that a peptide will be presented in a multiple-allele setting, are further discussed above in sections VIII.C.- VIII.C.7..

VIII.D.8. Training of Pan-Allele Models

[00431] Training a pan-allele model involves optimizing values for each parameter of the shared set of parameters $\boldsymbol{\theta}_H$ associated with the dependency function. Specifically, the parameters $\boldsymbol{\theta}_H$ are optimized such that the dependency function is able to output dependency scores that accurately indicate whether given MHC allele(s) will present a given peptide sequence.

[00432] To optimize the values of the parameters $\boldsymbol{\theta}_H$, the training data 170 is used. As mentioned above, the training data 170 used to train the model can include training samples that contain cells expressing single MHC alleles, training samples that contain cells expressing multiple MHC alleles, or training samples that contain cells expressing a combination of both single MHC alleles and multiple MHC alleles. Accordingly, each data instance i from the training data 170 is input into the pan-allele model, and more specifically, into the dependency function of the pan-allele model. For example, in certain embodiments, an MHC allele peptide sequence and a peptide sequence may be input into the pan-allele model. The pan-allele model then processes these inputs as if the model were being routinely used as described above with regard to sections VIII.D.3. – VIII.D.7.. However, unlike during the operation of the pan-allele model that is described in sections VIII.D.3. – VIII.D.7., during training of the pan-allele model, the known outcome of the peptide presentation is also input into the model. In other words, the label y^i is also input into the model. In embodiments in which the training sample input into the pan-allele model contains cells expressing multiple MHC alleles, y^i is set to 1 for each allele of the multiple MHC alleles in the sample.

[00433] After each iteration of the pan-allele model using a data instance i, the model determines the difference between the predicted probability of the MHC allele presenting the

peptide and the known label y^j. Then, to minimize this difference, the pan-allele model modifies the parameters θ_H. In other words, the pan-allele model determines values for the parameters θ_H by minimizing the loss function with respect to θ_H. When the pan-allele model achieves a certain level of prediction accuracy, the training is complete and the model is ready for use as described in sections VIII.D.3. – VIII.D.7..

VIII.D.9. Pan-Allele Model Examples

[00434] The following example compares the predictive precision (i.e. positive predictive value) of an example per-allele presentation model and an example pan-allele presentation model. In this example, the per-allele presentation model and the pan-allele presentation model are trained using the same training data set. Following training, the per-allele presentation model and the pan-allele presentation model are tested using six test samples. Note that the training data set contains ample training data for each MHC allele that is tested in each test sample. Table 2, below, shows the predictive precision (or positive predictive value) at a 40% recall rate when using the per-allele and the pan-allele model. Because of the ample training data for each MHC allele that is tested in the six samples, the per-allele model marginally outperforms the pan-allele model by 0.04 precision on average.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Mean
Pan-Allele	0.310	0.446	0.341	0.433	0.547	0.304	0.397
Per-Allele	0.363	0.517	0.349	0.458	0.565	0.368	0.437

Table 2

[00435] However, the ability of the pan-allele model to predict the presentation likelihood for an MHC allele that was *not* included in the training data set used to train the model can be observed in alternative experiments discussed with regard to FIGS. 16-22.

[00436] FIGS. 16-22 depict the results of experiments designed to test the ability of a pan-allele model to predict the probability that an untrained MHC allele will present a given peptide. In particular, FIGS. 16-18 depict the results of experiments designed to test the ability of a pan-allele model comprising a neural network model to predict the probability that an untrained MHC allele will present a given peptide. On the other hand, FIGS. 19-22 depict the results of experiments designed to test the ability of a pan-allele model comprising

a *non*-neural network model to predict the probability that an untrained MHC allele will present a given peptide.

[00437] Turning first to the experiments associated with FIGS. 16-18, to demonstrate the ability of a pan-allele model comprising a neural network model to predict the probability that an untrained MHC allele will present a given peptide, predictions generated by a pan-allele model comprising a neural network model that is *not* trained with the MHC alleles under test, are compared to predictions generated by an identical pan-allele model that *is* trained with the MHC alleles under test. In other words, the only difference between the pan-allele models is the set of training data on which they were trained. The greater the predictive precision of the pan-allele model that has been not trained on samples that include the tested HLA allele relative to the predictive precision of the pan-allele model that has been trained on samples that include the tested HLA allele, the greater the ability of a pan-allele model to predict presentation likelihood for MHC alleles that are not used to train the pan-allele model.

[00438] As noted above, the pan-allele models used within the experiments associated with FIGS. 16-18 are identical prior to training with the different training data sets. As also noted above, each of the pan-allele models used within the experiments associated with FIGS. 16-18 comprises a neural network model as its dependency function. The neural network model used in the pan-allele models contained a single hidden layer. The activation function between hidden layer of the neural network model was a rectified-linear unit (ReLU) function, $f(x) = \max(0, x)$. The last layer of the neural network model comprised a linear activation layer, $f(x) = x$. The number of hidden units per subnetwork of the neural network model was dependent on the inputs to the neural network model. Specifically, for neural network models configured to receive mRNA abundances, the number of hidden units in the mRNA abundance subnetwork of the neural network model was 16. For neural network models configured to receive encoded flanking sequences, the number of hidden units in the flanking sequence subnetwork of the neural network model was 32. For neural network models configured to receive encoded polypeptide sequences, the number of hidden units in the polypeptide sequence subnetwork of the neural network model was 256. For neural network models configured to receive encoded polypeptide sequences and encoded MHC allele peptide sequences (as in the case of the pan-allele models), the number of hidden units in the polypeptide and MHC allele peptide sequence subnetwork of the neural network model was 128.

[00439] Each experiment associated with FIGS. 16-18 includes a unique test sample, each unique test sample including a different HLA allele. To demonstrate that the results generated by these experiments are not restricted to a particular gene locus, an allele from each of the three gene loci, A, B, and C, was selected. Accordingly, the first test sample contains a HLA-A allele, the second sample contains a HLA-B allele, and the third sample contains a HLA-C allele. Specifically, the first test sample contains HLA allele A*02:03, the second test sample contains HLA allele B*54:01, and the third test sample contains HLA allele C*08:02. The protein sequence of each of these HLA alleles is obtained from the database of HLA protein sequences maintained by the Anthony Nolan Research Institute (https://www.ebi.ac.uk/ipd/imgt/hla/).

[00440] For each of the three samples, the protein sequence of the particular HLA allele and the protein sequence of the peptide in question are input into a first pan-allele model that has not been trained using the HLA allele, and into a second, identical pan-allele model that has been trained using the HLA allele. The pan-allele models output predicted probabilities that the HLA allele will present the peptide. These predicted probabilities are compared to the known outcome of the peptide presentation (i.e., the label y^i) to generate the precision/recall curves shown in FIGS. 16-18. Specifically, FIG. 16 corresponds to the data output by the pan-allele models for the first test sample, FIG. 17 corresponds to the data output by the pan-allele models for the second test sample, and FIG. 18 corresponds to the data output by the pan-allele models for the third test sample. In each figure, the blue line demonstrates the precision/recall curve for the pan-allele model that has been trained on samples that include the tested HLA allele, and the orange line demonstrates the precision/recall curve for the pan-allele model that has not been trained on any samples that include the tested HLA allele. Additionally, each figure indicates the average predictive precision (i.e., positive predictive value) of both the trained and untrained pan-allele models. For example, as seen in FIG. 18, the average predictive precision of the pan-allele model that has been trained on samples that include the tested HLA allele is 0.256 and the average predictive precision of the pan-allele model that has not been trained on samples that include the tested HLA allele is 0.231.

[00441] As shown in FIGS. 16-18, even though the pan-allele models represented by the orange lines have never seen the HLA allele under test, these pan allele models are able to achieve comparable performance to the pan-allele models represented by the blue lines that have seen the HLA allele under test during training. Therefore, these results demonstrate the

ability of a pan-allele model comprising a neural network model, to accurately predict presentation likelihoods for HLA alleles that were not used to train the pan-allele model.

[00442] Turning next to the experiments associated with FIGS. 19-22, to demonstrate the ability of a pan-allele model comprising a non-neural network model to predict the probability that an untrained MHC allele will present a given peptide, the performance of four models are compared within each experiment. The four models include: a pan-allele presentation model comprising a neural network model as described above with regard to FIGS. 16-18, an off-the-shelf random forest model composed of 1,000 trees, an off-the-shelf quadratic discriminant analysis (QDA) model that fits multivariate Gaussians, and a current state-of-the-art MHC class 1 binding affinity model MHCFlurry that fits a distinct feed-forward, fully-connected neural network for each allele. The random forest model and the quadratic discriminant model are both based on pan-allele model architecture that comprises a non-neural network model.

[00443] Each experiment associated with FIGS. 19-22 includes a test sample, and each test sample includes an HLA allele. To demonstrate that the results generated by these experiments are not restricted to a particular gene locus, an allele from each of the three gene loci, A, B, and C, was selected. Accordingly, a first test sample and a second test sample contain a HLA-A allele, a third sample contains a HLA-B allele, and a fourth sample contains a HLA-C allele. Specifically, the first test sample and the second test sample contain HLA allele A*02:01, the third test sample contains HLA allele B*44:02, and the fourth test sample contains HLA allele C*08:02. The protein sequence of each of these HLA alleles is obtained from the database of HLA protein sequences maintained by the Anthony Nolan Research Institute (https://www.ebi.ac.uk/ipd/imgt/hla/).

[00444] During training of the four models used to predict presentation likelihoods for each of the four test samples, the pan-allele presentation model, the random forest model, and the quadratic discriminant model are each trained on single-allele data composed of 9-mers from 31 distinct alleles and including HLA-A, HLA-B, and HLA-C. On the other hand, the MHCFlurry model is trained by its authors using a subset of the IEDB and BD2013 binding affinity data sets, including alleles from HLA-A, HLA-B, and HLA-C. Each allele is modeled individually with an ensemble of 8 neural networks, and the allele name is directly passed to the model to select which allele-submodel to use to generate presentation prediction. [76].

[00445] The particular alleles used to train the four models for each of the four test samples are dependent upon the HLA allele contained within the given test sample. Specifically, for the first test sample that contains HLA allele A*02:01, the training data used to train the four models to predict a presentation likelihood for the HLA allele A*02:01, includes the HLA allele A*02:01. For the second test sample that contains HLA allele A*02:01, the training data used to train the four models to predict a presentation likelihood for the HLA allele A*02:01, does *not* include the HLA allele A*02:01. For the third test sample that contains HLA allele B*44:02, the training data used to train the four models to predict a presentation likelihood for the HLA allele B*44:02, does *not* include the HLA allele B*44:02. For the fourth test sample that contains HLA allele C*08:02, the training data used to train the four models to predict a presentation likelihood for the HLA allele C*08:02, does *not* include the HLA allele C*08:02.

[00446] During testing for each of the four samples, each model was tested on a held-out single-allele dataset comprising the HLA allele in the given sample, and composed of about 250,000 peptides (counting both presented and non-presented peptides). Specifically, during testing for each of the four samples, the pan-allele presentation model, the random forest model, and the quadratic discriminant model each received the same input. Particularly, for each of the four samples, the pan-allele presentation model, the random forest model, and the quadratic discriminant model each received the 34-mer one-hot encoded HLA allele protein sequence of the HLA allele within the sample, and the 9-mer one-hot encoded (i.e., binarized) protein sequence of the peptide in question. On the other hand, for each of the four samples, the MHCFlurry model received the name of the HLA allele within the sample, and the 9-mer one hot encoded (i.e., binarized) protein sequence of the peptide in question. As described above, this discrepancy in inputs between the models is a result of the fact that the MHCFlurry model is configured to use the name of an allele to select which allele-submodel to use to generate a presentation prediction.

[00447] Following these inputs into the four models, each of the four models then outputs a predicted probability that the HLA allele will present the peptide. These predicted probabilities are compared to the known outcome of the peptide presentation (i.e., the label y^i) to generate the precision/recall curves shown in FIGS. 19-22. Specifically, FIG. 19 corresponds to the data output by each of the four models for the first test sample, FIG. 20 corresponds to the data output by each of the four models for the second test sample, FIG. 21 corresponds to the data output by each of the four models for the third test sample, and FIG.

22 corresponds to the data output by each of the four models for the fourth test sample. In each figure, the blue line demonstrates the precision/recall curve for the pan-allele model, the orange line demonstrates the precision/recall curve for the MHCFlurry model, the green line demonstrates the precision/recall curve for the random forest model, and the red line demonstrates the precision/recall curve for the quadratic discriminant model. Additionally, each figure indicates the average predictive precision (i.e., positive predictive value) of each of the models. For example, as seen in FIG. 19, the average predictive precision of the pan-allele model is 0.32.

[00448] As shown in FIGS. 19-22, the random forest model and the quadratic discriminant model that both used the pan-allele model architecture comprising a non-neural network model, both performed about twice as well as the MHCFlurry model. Furthermore, the pan-allele presentation model comprising the neural network model performed about twice as well as the random forest model and the quadratic discriminant model that used the pan-allele model architecture comprising the non-neural network model. In other words, the pan-allele presentation model comprising the neural network model achieved the highest precision relative to the other models. However, the random forest model and the quadratic discriminant model that used the pan-allele model architecture comprising the non-neural network model still outperformed the custom-made per-allele binding affinity model MHCFlurry. Therefore, these results demonstrate that the pan-allele model architecture can generalize well to other non-neural network machine learning models that are as varied as decision-tree based random forests and Bayesian methods like quadratic discriminant analysis, while still providing high levels of predictive precision.

[00449] Additionally, as further shown in FIGS. 20-22, even though the pan-allele presentation model, the random forest model, and the quadratic discriminant model have never seen the HLA allele under test, these models, including the random forest model and the quadratic discriminant model that both used the pan-allele model architecture comprising a non-neural network model, are able to achieve comparable performance to the models corresponding to FIG. 19 that have seen the HLA allele under test during training. Therefore, these results demonstrate the ability of the pan-allele model architecture that comprises a non-neural network to accurately predict presentation likelihoods for HLA alleles that were not used to train the model.

IX. Example 5: Prediction Module

[00450] The prediction module 320 receives sequence data and selects candidate neoantigens in the sequence data using the presentation models. Specifically, the sequence data may be DNA sequences, RNA sequences, and/or protein sequences extracted from tumor tissue cells of patients. The prediction module 320 processes the sequence data into a plurality of peptide sequences p^k having 8-15 amino acids for MHC-I or 6-30 amino acids for MHC-II. For example, the prediction module 320 may process the given sequence "IEFROEIFJEF into three peptide sequences having 9 amino acids "IEFROEIFJ," "EFROEIFJE," and "FROEIFJEF." In one embodiment, the prediction module 320 may identify candidate neoantigens that are mutated peptide sequences by comparing sequence data extracted from normal tissue cells of a patient with the sequence data extracted from tumor tissue cells of the patient to identify portions containing one or more mutations.

[00451] The prediction module 320 applies one or more of the presentation models to the processed peptide sequences to estimate presentation likelihoods of the peptide sequences. Specifically, the prediction module 320 may select one or more candidate neoantigen peptide sequences that are likely to be presented on tumor HLA molecules by applying the presentation models to the candidate neoantigens. In one implementation, the prediction module 320 selects candidate neoantigen sequences that have estimated presentation likelihoods above a predetermined threshold. In another implementation, the presentation model selects the v candidate neoantigen sequences that have the highest estimated presentation likelihoods (where v is generally the maximum number of epitopes that can be delivered in a vaccine). A vaccine including the selected candidate neoantigens for a given patient can be injected into the patient to induce immune responses.

X. Example 6: Patient Selection Module

[00452] The patient selection module 324 selects a subset of patients for vaccine treatment and/or T-cell therapy based on whether the patients satisfy inclusion criteria. In one embodiment, the inclusion criteria is determined based on the presentation likelihoods of patient neoantigen candidates as generated by the presentation models. By adjusting the inclusion criteria, the patient selection module 324 can adjust the number of patients that will receive the vaccine and/or T-cell therapy based on his or her presentation likelihoods of neoantigen candidates. Specifically, a stringent inclusion criteria results in a fewer number of patients that will be treated with the vaccine and/or T-cell therapy, but may result in a higher proportion of vaccine and/or T-cell therapy-treated patients that receive effective treatment

(e.g., 1 or more tumor-specific neoantigens (TSNA) and/or 1 or more neoantigen-responsive T-cells). On the other hand, a lenient inclusion criteria results in a higher number of patients that will be treated with the vaccine and/or with T-cell therapy, but may result in a lower proportion of vaccine and/or T-cell therapy-treated patients that receive effective treatment. The patient selection module 324 modifies the inclusion criteria based on the desired balance between target proportion of patients that will receive treatment and proportion of patients that receive effective treatment.

[00453] In some embodiments, inclusion criteria for selection of patients to receive vaccine treatment are the same as inclusion criteria for selection of patients to receive T-cell therapy. However, in alternative embodiments, inclusion criteria for selection of patients to receive vaccine treatment may differ from inclusion criteria for selection of patients to receive T-cell therapy. The following Sections X.A and X.B discuss inclusion criteria for selection of patients to receive vaccine treatment and inclusion criteria for selection of patients to receive T-cell therapy, respectively.

X.A. Patient Selection for Vaccine Treatment

[00454] In one embodiment, patients are associated with a corresponding treatment subset of v neoantigen candidates that can potentially be included in customized vaccines for the patients with vaccine capacity v. In one embodiment, the treatment subset for a patient are the neoantigen candidates with the highest presentation likelihoods as determined by the presentation models. For example, if a vaccine can include v=20 epitopes, the vaccine can include the treatment subset of each patient that have the highest presentation likelihoods as determined by the presentation model. However, it is appreciated that in other embodiments, the treatment subset for a patient can be determined based on other methods. For example, the treatment subset for a patient may be randomly selected from the set of neoantigen candidates for the patient, or may be determined in part based on current state-of-the-art models that model binding affinity or stability of peptide sequences, or some combination of factors that include presentation likelihoods from the presentation models and affinity or stability information regarding those peptide sequences.

[00455] In one embodiment, the patient selection module 324 determines that a patient satisfies the inclusion criteria if the tumor mutation burden of the patient is equal to or above a minimum mutation burden. The tumor mutation burden (TMB) of a patient indicates the total number of nonsynonymous mutations in the tumor exome. In one implementation, the patient selection module 324 may select a patient for vaccine treatment if the absolute

number of TMB of the patient is equal to or above a predetermined threshold. In another implementation, the patient selection module 324 may select a patient for vaccine treatment if the TMB of the patient is within a threshold percentile among the TMB's determined for the set of patients.

[00456] In another embodiment, the patient selection module 324 determines that a patient satisfies the inclusion criteria if a utility score of the patient based on the treatment subset of the patient is equal to or above a minimum utility score. In one implementation, the utility score is a measure of the estimated number of presented neoantigens from the treatment subset.

[00457] The estimated number of presented neoantigens may be predicted by modeling neoantigen presentation as a random variable of one or more probability distributions. In one implementation, the utility score for patient i is the expected number of presented neoantigen candidates from the treatment subset, or some function thereof. As an example, the presentation of each neoantigen can be modeled as a Bernoulli random variable, in which the probability of presentation (success) is given by the presentation likelihood of the neoantigen candidate. Specifically, for a treatment subset S_i of v neoantigen candidates $\boldsymbol{p}^{i1}$, $\boldsymbol{p}^{i2}$, ..., $\boldsymbol{p}^{iv}$ each having the highest presentation likelihoods u_{i1}, u_{i2}, ..., u_{iv}, presentation of neoantigen candidate $\boldsymbol{p}^{ij}$ is given by random variable A_{ij}, in which:

$$P(A_{ij} = 1) = u_{ij}, \qquad P(A_{ij} = 0) = 1 - u_{ij}. \tag{29}$$

The expected number of presented neoantigens is given by the summation of the presentation likelihoods for each neoantigen candidate. In other words, the utility score for patient i can be expressed as:

$$\text{util}_i(S_i) = \mathbb{E}\left[\sum_{j=1}^{v} A_{ij}\right] = \sum_{j=1}^{v} u_{ij}. \tag{30}$$

The patient selection module 324 selects a subset of patients having utility scores equal to or above a minimum utility for vaccine treatment.

[00458] In another implementation, the utility score for patient i is the probability that at least a threshold number of neoantigens k will be presented. In one instance, the number of presented neoantigens in the treatment subset S_i of neoantigen candidates is modeled as a Poisson Binomial random variable, in which the probabilities of presentation (successes) are given by the presentation likelihoods of each of the epitopes. Specifically, the number of presented neoantigens for patient i can be given by random variable N_i, in which:

$$N_i = \sum_{j=1}^{v} A_{ij} \sim \text{PBD}(u_{i1}, u_{i2}, \ldots, u_{iv}). \quad (31)$$

where PBD(·) denotes the Poisson Binomial distribution. The probability that at least a threshold number of neoantigens k will be presented is given by the summation of the probabilities that the number of presented neoantigens N_i will be equal to or above k. In other words, the utility score for patient i can be expressed as:

$$\text{util}_i(S_i) = \mathbb{P}[N_i \geq k] = \sum_{m=1}^{k} \mathbb{P}[N_i = m]. \quad (32)$$

The patient selection module 324 selects a subset of patients having the utility score equal to or above a minimum utility for vaccine treatment.

[00459] In another implementation, the utility score for patient i is the number of neoantigens in the treatment subset S_i of neoantigen candidates having binding affinity or predicted binding affinity below a fixed threshold (e.g., 500nM) to one or more of the patient's HLA alleles. In one instance, the fixed threshold is a range from 1000nM to 10nM. Optionally, the utility score may count only those neoantigens detected as expressed via RNA-seq.

[00460] In another implementation, the utility score for patient i is the number of neoantigens in the treatment subset S_i of neoantigen candidates having binding affinity to one or more of that patient's HLA alleles at or below a threshold percentile of binding affinities for random peptides to that HLA allele. In one instance, the threshold percentile is a range from the 10th percentile to the 0.1^{th} percentile. Optionally, the utility score may count only those neoantigens detected as expressed via RNA-seq.

[00461] It is appreciated that the examples of generating utility scores illustrated with respect to equations (25) and (27) are merely illustrative, and the patient selection module 324 may use other statistics or probability distributions to generate the utility scores.

X.B. Patient Selection for T-Cell Therapy

[00462] In another embodiment, instead of or in addition to receiving vaccine treatment, patients can receive T-cell therapy. Like vaccine treatment, in embodiments in which a patient receives T-cell therapy, the patient may be associated with a corresponding treatment subset of v neoantigen candidates as described above. This treatment subset of v neoantigen candidates can be used for *in vitro* identification of T cells from the patient that are

responsive to one or more of the *v* neoantigen candidates. These identified T cells can then be expanded and infused into the patient for customized T-cell therapy.

[00463] Patients may be selected to receive T-cell therapy at two different time points. The first point is after the treatment subset of *v* neoantigen candidates have been predicted for a patient using the models, but before *in vitro* screening for T cells that are specific to the predicted treatment subset of *v* neoantigen candidates. The second point is after *in vitro* screening for T cells that are specific to the predicted treatment subset of *v* neoantigen candidates.

[00464] First, patients may be selected to receive T-cell therapy after the treatment subset of *v* neoantigen candidates have been predicted for the patient, but before *in vitro* identification of T-cells from the patient that are specific to the predicted subset of *v* neoantigen candidates. Specifically, because *in vitro* screening for neoantigen-specific T-cells from the patient can be expensive, it may be desirable to only select patients to screen for neoantigen-specific T-cells if the patients are likely to have neoantigen-specific T-cells. To select patients before the *in vitro* T-cell screening step, the same criteria that are used to select patients for vaccine treatment may be used. Specifically, in some embodiments, the patient selection module 324 may select a patient to receive T-cell therapy if the tumor mutation burden of the patient is equal to or above a minimum mutation burden as described above. In another embodiment, the patient selection module 324 may select a patient to receive T-cell therapy if a utility score of the patient based on the treatment subset of *v* neoantigen candidates for the patient is equal to or above a minimum utility score, as described above.

[00465] Second, in addition to or instead of selecting patients to receive T-cell therapy before *in vitro* identification of T-cells from the patient that are specific to the predicted subset of *v* neoantigen candidates, patients may also be selected to receive T-cell therapy after *in vitro* identification of T-cells that are specific to the predicted treatment subset of *v* neoantigen candidates. Specifically, a patient may be selected to receive T-cell therapy if at least a threshold quantity of neoantigen-specific TCRs are identified for the patient during the *in vitro* screening of the patient's T-cells for neoantigen recognition. For example, a patient may be selected to receive T-cell therapy only if at least two neoantigen-specific TCRs are identified for the patient, or only if neoantigen-specific TCRs are identified for two distinct neoantigens.

[00466] In another embodiment, a patient may be selected to receive T-cell therapy only if at least a threshold quantity of neoantigens of the treatment subset of *v* neoantigen candidates for the patient are recognized by the patient's TCRs. For example, a patient may be selected to receive T-cell therapy only if at least one neoantigen of the treatment subset of *v* neoantigen candidates for the patient are recognized by the patient's TCRs. In further embodiments, a patient may be selected to receive T-cell therapy only if at least a threshold quantity of TCRs for the patient are identified as neoantigen-specific to neoantigen peptides of a particular HLA restriction class. For example, a patient may be selected to receive T-cell therapy only if at least one TCR for the patient is identified as neoantigen-specific HLA class I restricted neoantigen peptides.

[00467] In even further embodiments, a patient may be selected to receive T-cell therapy only if at least a threshold quantity of neoantigen peptides of a particular HLA restriction class are recognized by the patient's TCRs. For example, a patient may be selected to receive T-cell therapy only if at least one HLA class I restricted neoantigen peptide is recognized by the patient's TCRs. As another example, a patient may be selected to receive T-cell therapy only if at least two HLA class II restricted neoantigen peptides are recognized by the patient's TCRs. Any combination of the above criteria may also be used for selecting patients to receive T-cell therapy after *in vitro* identification of T-cells that are specific to the predicted treatment subset of *v* neoantigen candidates for the patient.

XI. Example 7: Experimentation Results Showing Example Patient Selection Performance

[00468] The validity of patient selection methods described in Section X are tested by performing patient selection on a set of simulated patients each associated with a test set of simulated neoantigen candidates, in which a subset of simulated neoantigens is known to be presented in mass spectrometry data. Specifically, each simulated neoantigen candidate in the test set is associated with a label indicating whether the neoantigen was presented in a multiple-allele JY cell line HLA-A*02:01 and HLA-B*07:02 mass spectrometry data set from the Bassani-Sternberg data set (data set "D1") (data can be found at www.ebi.ac.uk/pride/archive/projects/PXD0000394). As described in more detail below in conjunction with FIG. 23A, a number of neoantigen candidates for the simulated patients are sampled from the human proteome based on the known frequency distribution of mutation burden in non-small cell lung cancer (NSCLC) patients.

[00469] Per-allele presentation models for the same HLA alleles are trained using a training set that is a subset of the single-allele HLA-A*02:01 and HLA-B*07:02 mass spectrometry data from the IEDB data set (data set "D2") (data can be found at http://www.iedb.org/doc/mhc_ligand_full.zip). Specifically, the presentation model for each allele was the per-allele model shown in equation (8) that incorporated N-terminal and C-terminal flanking sequences as allele-noninteracting variables, with network dependency functions $g_h(\cdot)$ and $g_w(\cdot)$, and the expit function $f(\cdot)$. The presentation model for allele HLA-A*02:01 generates a presentation likelihood that a given peptide will be presented on allele HLA-A*02:01, given the peptide sequence as an allele-interacting variable, and the N-terminal and C-terminal flanking sequences as allele-noninteracting variables. The presentation model for allele HLA-B*07:02 generates a presentation likelihood that a given peptide will be presented on allele HLA-B*07:02, given the peptide sequence as an allele-interacting variable, and the N-terminal and C-terminal flanking sequences as allele-noninteracting variables.

[00470] As laid out in the following examples and with reference to FIGS. 23A-23E, various models, such as the trained presentation models and current state-of-the-art models for peptide binding prediction, are applied to the test set of neoantigen candidates for each simulated patient to identify different treatment subsets for patients based on the predictions. Patients that satisfy inclusion criteria are selected for vaccine treatment, and are associated with customized vaccines that include epitopes in the treatment subsets of the patients. The size of the treatment subsets are varied according to different vaccine capacities. No overlap is introduced between the training set used to train the presentation model and the test set of simulated neoantigen candidates.

[00471] In the following examples, the proportion of selected patients having at least a certain number of presented neoantigens among the epitopes included in the vaccines are analyzed. This statistic indicates the effectiveness of the simulated vaccines to deliver potential neoantigens that will elicit immune responses in patients. Specifically, a simulated neoantigen in a test set is presented if the neoantigen is presented in the mass spectrometry data set D2. A high proportion of patients with presented neoantigens indicate potential for successful treatment via neoantigen vaccines by inducing immune responses.

XI.A. Example 7A: Frequency Distribution of Mutation Burden for NSCLC Cancer Patients

[00472] FIG. 23A illustrates a sample frequency distribution of mutation burden in NSCLC patients. Mutation burden and mutations in different tumor types, including NSCLC, can be found, for example, at the cancer genome atlas (TCGA) (https://cancergenome.nih.gov). The x-axis represents the number of non-synonymous mutations in each patient, and the y-axis represents the proportion of sample patients that have the given number of non-synonymous mutations. The sample frequency distribution in FIG. 23A shows a range of 3-1786 mutations, in which 30% of the patients have fewer than 100 mutations. Although not shown in FIG. 23A, research indicates that mutation burden is higher in smokers compared to that of non-smokers, and that mutation burden may be a strong indicator of neoantigen load in patients.

[00473] As introduced at the beginning of Section XI above, each of a number of simulated patients are associated with a test set of neoantigen candidates. The test set for each patient is generated by sampling a mutation burden m_i from the frequency distribution shown in FIG. 23A for each patient. For each mutation, a 21-mer peptide sequence from the human proteome is randomly selected to represent a simulated mutated sequence. A test set of neoantigen candidate sequences are generated for patient i by identifying each (8, 9, 10, 11)-mer peptide sequence spanning the mutation in the 21-mer. Each neoantigen candidate is associated with a label indicating whether the neoantigen candidate sequence was present in the mass spectrometry D1 data set. For example, neoantigen candidate sequences present in data set D1 may be associated with a label "1," while sequences not present in data set D1 may be associated with a label "0." As described in more detail below, FIGS. 23B through 23E illustrate experimental results on patient selection based on presented neoantigens of the patients in the test set.

XI.B. Example 7B: Proportion of Selected Patients with Neoantigen Presentation based on Mutation Burden Inclusion Criteria

[00474] FIG. 23B illustrates the number of presented neoantigens in simulated vaccines for patients selected based on an inclusion criteria of whether the patients satisfy a minimum mutation burden. The proportion of selected patients that have at least a certain number of presented neoantigens in the corresponding test is identified.

[00475] In FIG. 23B, the x-axis indicates the proportion of patients excluded from vaccine treatment based on the minimum mutation burden, as indicated by the label "minimum # of

mutations." For example, a data point at 200 "minimum # of mutations" indicates that the patient selection module 324 selected only the subset of simulated patients having a mutation burden of at least 200 mutations. As another example, a data point at 300 "minimum # of mutations" indicates that the patient selection module 324 selected a lower proportion of simulated patients having at least 300 mutations. The y-axis indicates the proportion of selected patients that are associated with at least a certain number of presented neoantigens in the test set without any vaccine capacity v. Specifically, the top plot shows the proportion of selected patients that present at least 1 neoantigen, the middle plot shows the proportion of selected patients that present at least 2 neoantigens, and the bottom plot shows the proportion of selected patients that present at least 3 neoantigens.

[00476] As indicated in FIG. 23B, the proportion of selected patients with presented neoantigens increases significantly with higher mutation burden. This indicates that mutation burden as an inclusion criteria can be effective in selecting patients for whom neoantigen vaccines are more likely to induce successful immune responses.

XI.C. Example 7C: Comparison of Neoantigen Presentation for Vaccines Identified by Presentation Models vs. State-of-the-Art Models

[00477] FIG. 23C compares the number of presented neoantigens in simulated vaccines between selected patients associated with vaccines including treatment subsets identified based on presentation models and selected patients associated with vaccines including treatment subsets identified through current state-of-the-art models. The left plot assumes limited vaccine capacity v=10, and the right plot assumes limited vaccine capacity v=20. The patients are selected based on utility scores indicating expected number of presented neoantigens.

[00478] In FIG. 23C, the solid lines indicate patients associated with vaccines including treatment subsets identified based on presentation models for alleles HLA-A*02:01 and HLA-B*07:02. The treatment subset for each patient is identified by applying each of the presentation models to the sequences in the test set, and identifying the v neoantigen candidates that have the highest presentation likelihoods. The dotted lines indicate patients associated with vaccines including treatment subsets identified based on current state-of-the-art models NETMHCpan for the single allele HLA-A*02:01. Implementation details for NETMHCpan is provided in detail at http://www.cbs.dtu.dk/services/NetMHCpan. The treatment subset for each patient is identified by applying the NETMHCpan model to the sequences in the test set, and identifying the v neoantigen candidates that have the highest

estimated binding affinities. The x-axis of both plots indicates the proportion of patients excluded from vaccine treatment based on expectation utility scores indicating the expected number of presented neoantigens in treatment subsets identified based on presentation models. The expectation utility score is determined as described in reference to equation (25) in Section X. The y-axis indicates the proportion of selected patients that present at least a certain number of neoantigens (1, 2, or 3 neoantigens) included in the vaccine.

[00479] As indicated in FIG. 23C, patients associated with vaccines including treatment subsets based on presentation models receive vaccines containing presented neoantigens at a significantly higher rate than patients associated with vaccines including treatment subsets based on state-of-the-art models. For example, as shown in the right plot, 80% of selected patients associated with vaccines based on presentation models receive at least one presented neoantigen in the vaccine, compared to only 40% of selected patients associated with vaccines based on current state-of-the-art models. The results indicate that presentation models as described herein are effective for selecting neoantigen candidates for vaccines that are likely to elicit immune responses for treating tumors.

XI.D. Example 7D: Effect of HLA Coverage on Neoantigen Presentation for Vaccines Identified Through Presentation Models

[00480] FIG. 23D compares the number of presented neoantigens in simulated vaccines between selected patients associated with vaccines including treatment subsets identified based on a single per-allele presentation model for HLA-A*02:01 and selected patients associated with vaccines including treatment subsets identified based on both per-allele presentation models for HLA-A*02:01 and HLA-B*07:02. The vaccine capacity is set as v=20 epitopes. For each experiment, the patients are selected based on expectation utility scores determined based on the different treatment subsets.

[00481] In FIG. 23D, the solid lines indicate patients associated with vaccines including treatment subsets based on both presentation models for HLA alleles HLA-A*02:01 and HLA-B*07:02. The treatment subset for each patient is identified by applying each of the presentation models to the sequences in the test set, and identifying the v neoantigen candidates that have the highest presentation likelihoods. The dotted lines indicate patients associated with vaccines including treatment subsets based on a single presentation model for HLA allele HLA-A*02:01. The treatment subset for each patient is identified by applying the presentation model for only the single HLA allele to the sequences in the test set, and identifying the v neoantigen candidates that have the highest presentation likelihoods. For

solid line plots, the x-axis indicates the proportion of patients excluded from vaccine treatment based on expectation utility scores for treatment subsets identified by both presentation models. For dotted line plots, the x-axis indicates the proportion of patients excluded from vaccine treatment based on expectation utility scores for treatment subsets identified by the single presentation model. The y-axis indicates the proportion of selected patients that present at least a certain number of neoantigens (1, 2, or 3 neoantigens).

[00482] As indicated in FIG. 23D, patients associated with vaccines including treatment subsets identified by presentation models for both HLA alleles present neoantigens at a significantly higher rate than patients associated with vaccines including treatment subsets identified by a single presentation model. The results indicate the importance of establishing presentation models with high HLA allele coverage.

XI.E. Example 7E: Comparison of Neoantigen Presentation for Patients Selected by Mutation Burden vs. Expected Number of Presented Neoantigens

[00483] FIG. 23E compares the number of presented neoantigens in simulated vaccines between patients selected based on mutation burden and patients selected by expectation utility score. The expectation utility scores are determined based on treatment subsets identified by presentation models having a size of $v=20$ epitopes.

[00484] In FIG. 23E, the solid lines indicate patients selected based on expectation utility score associated with vaccines including treatment subsets identified by presentation models. The treatment subset for each patient is identified by applying the presentation models to sequences in the test set, and identifying the $v=20$ neoantigen candidates that have the highest presentation likelihoods. The expectation utility score is determined based on the presentation likelihoods of the identified treatment subset based on equation (25) in section X. The dotted lines indicate patients selected based on mutation burden associated with vaccines also including treatment subsets identified by presentation models. The x-axis indicates the proportion of patients excluded from vaccine treatment based on expectation utility scores for solid line plots, and proportion of patients excluded based on mutation burden for dotted line plots. The y-axis indicates the proportion of selected patients who receive a vaccine containing at least a certain number of presented neoantigens (1, 2, or 3 neoantigens).

As indicated in FIG. 23E, patients selected based on expectation utility scores receive a vaccine containing presented neoantigens at a higher rate than patients selected based on

mutation burden. However, patients selected based on mutation burden receive a vaccine containing presented neoantigens at a higher rate than unselected patients. Thus, mutation burden is an effective patient selection criteria for successful neoantigen vaccine treatment, though expectation utility scores are more effective.

XII. Example 8: Evalution of Mass Spectrometry-Trained Model on Held-Out Mass Spectrometry Data

[00485] As HLA peptide presentation by tumor cells is a key requirement for anti-tumor immunity[91,96,97], a large (N=74 patients) integrated dataset of human tumor and normal tissue samples with paired class I HLA peptide sequences, HLA types and transcriptome RNA-seq (Methods) was generated with the aim of using these and publicly available data[92,98,99] to train a novel deep learning model[100] to predict antigen presentation in human cancer. Samples were chosen among several tumor types of interest for immunotherapy development and based on tissue availability. Mass spectrometry identified an average of 3,704 peptides per sample at peptide-level FDR<0.1 (range 344-11,301). The peptides followed the characteristic class I HLA length distribution: lengths 8-15aa, with a modal length of 9 (56% of peptides). Consistent with previous reports, a majority of peptides (median 79%) were predicted to bind at least one patient HLA allele at the standard 500nM affinity threshold by MHCflurry[90], but with substantial variability across samples (e.g., 33% of peptides in one sample had predicted affinities >500nM). The commonly used[101] "strong binder" threshold of 50nM captured a median of only 42% of presented peptides. Transcriptome sequencing yielded an average of 131M unique reads per sample and 68% of genes were expressed at a level of at least 1 transcript per million (TPM) in at least one sample, highlighting the value of a large and diverse sample set to observe expression of a maximal number of genes. Peptide presentation by the HLA was strongly correlated with mRNA expression. Striking and reproducible gene-to-gene differences in the rate of peptide presentation, beyond what could be explained by differences in RNA expression or sequence alone, were observed. The observed HLA types matched expectations for specimens from a predominantly European-ancestry group of patients.

[00486] For each patient, the positive-labeled data points were peptides detected via mass spectrometry, and the negative-labeled data points were peptides from the reference proteome (SwissProt) that were not detected via mass spectrometry in that sample. The data was split into training, validation and testing sets (Methods). The training set consisted of 142,844

HLA presented peptides (FDR<~0.02) from 101 samples (69 newly described in this study and 32 previously published). The validation set (used for early stopping) consisted of 18,004 presented peptides from the same 101 samples. Two mass spectrometry datasets were used for testing: (1) A tumor sample test set consisting of 571 presented peptides from 5 additional tumor samples (2 lung, 2 colon, 1 ovarian) that were held out of the training data, and (2) a single-allele cell line test set consisting of 2,128 presented peptides from genomic location windows (blocks) adjacent to, but distinct from, the locations of single-allele peptides included in the training data (see Methods for additional details on the train/test splits).

[00487] Using these and publicly available HLA peptide data[92,98,99], neural network (NN) models were trained to predict HLA antigen presentation. Specifically, in Example 9, the pan-allele model discussed above in Section VIII.D was trained using the above data to predict HLA antigen presentation. On the other hand, in Example 11, an allele-specific model described in detail below was trained using the above data to predict HLA antigen presentation. In Example 10, both the pan-allele model discussed above in Section VIII.D and the allele-specific model described in detail below were trained using the above data to predict HLA antigen presentation.

[00488] In particular, in Examples 10 and 11, to learn allele-specific models from tumor mass spectrometry data where each peptide could have been presented by any one of six HLA alleles, a novel network architecture capable of jointly learning the allele-peptide mappings and allele-specific presentation motifs (*see* Section XVII.B below) was developed. The training data identified predictive models for 53 HLA alleles. In contrast to prior work[92,104], these models captured the dependence of HLA presentation on each sequence position for peptides of multiple lengths. The model also correctly learned the critical dependencies on gene RNA expression and gene-specific presentation propensity, with the mRNA abundance and learned per-gene propensity of presentation combining independently to yield up to a ~60-fold difference in rate of presentation between the lowest-expressed, least presentation-prone and the highest expressed, most presentation-prone genes. It was further observed that the model predicted the measured stability of HLA/peptide complexes in IEDB[88] ($p<1e-10$ for 10 alleles), even after controlling for predicted binding affinity ($p<0.05$ for 8/10 alleles tested). Collectively, these features form the basis for improved prediction of immunogenic HLA class I peptides.

XIII. Example 9: Experimentation Results Including Presentation Hotspot Modeling

[00489] To specifically evaluate the benefit of using presentation hotspot parameters in modeling HLA presentation, the performance of a pan-allele neural network presentation model that incorporates presentation hotspot parameters was compared with the performance of a pan-allele neural network presentation model that does not incorporate presentation hotspot parameters. The base neural network architecture was the same for both pan-allele models and was identical to the pan-allele presentation model described above in Sections VII-VIII. In brief, the pan-allele models included peptide and flanking amino acid sequence parameters, RNA-sequencing transcription data (TPM), protein family data, per-sample identification, and HLA-A, B, C types. Ensembles of 5 networks were used for each pan-allele model. The pan-allele model that included the presentation hotspot parameters used Equation 12b described above in Section VIII.B.3., with a per-gene proteomic block size of 10, and peptide lengths 8-12.

[00490] The two pan-allele models were compared by performing experiments using the mass spectrometry dataset described above in Section XII. Specifically, five samples were held-out from model training and validation for the purpose of fairly evaluating the competing models. The remaining samples were randomly divided 90% for model training and 10% for validating the training.

[00491] FIG. 24 compares the positive predictive values (PPVs) at 40% recall of a pan-allele presentation model that uses presentation hotspot parameters and a pan-allele presentation model that does not use presentation hotspot parameters, when the pan-allele models are tested on five held-out test samples. As shown in FIG. 24, the pan-allele presentation model that incorporated presentation hotspot parameters consistently out-performed the pan-allele presentation model that did not incorporate presentation hotspot parameters.

XIV. Example 10: Model Evaluation of Retrospective Neoantigen T-Cell Data

[00492] We then evaluated whether the accurate prediction of HLA peptide presentation of the pan-allele model could translate into the ability to identify human tumor CD8 T-cell epitopes (i.e., immunotherapy targets). Defining an appropriate test dataset for this evaluation is challenging, as it requires peptides that are both recognized by T-cells and presented by the HLA on the tumor cell surface. In addition, formal performance assessment requires not only positive-labeled (i.e., T-cell recognized) peptides, but also a sufficient number of negative-

labeled (i.e., tested but not recognized) peptides. Mass spectrometry datasets address tumor presentation but not T-cell recognition; oppositely, priming or T-cell assays post-vaccination address the presence of T-cell precursors and T-cell recognition but not tumor presentation (for example, a strong-binding peptide whose source gene is expressed in the tumor at too low a level to support presentation of the peptide could give rise to a strong CD8 T-cell response after administration of a vaccine but would not be a therapeutically useful target, because it is not presented by the tumor).

[00493] To obtain an appropriate dataset, we collected published CD8 T-cell epitopes from 4 recent studies that met the required criteria: study A[140] examined TIL in 9 patients with gastrointestinal tumors and reported T-cell recognition of 12/1,053 somatic SNV mutations tested by IFN-γ ELISPOT using the tandem minigene (TMG) method in autologous DCs. Study B[84] also used TMGs and reported T-cell recognition of 6/574 SNVs by CD8+PD-1+ circulating lymphocytes from 4 melanoma patients. Study C[141] assessed TIL from 3 melanoma patients using pulsed peptide stimulation and found responses to 5/381 tested SNV mutations. Study D[108] assessed TIL from one breast cancer patient using a combination of TMG assays and pulsing with minimal epitope peptides and reported recognition of 2/62 SNVs. The combined dataset consisted of 2,023 assayed SNVs from 17 patients, including 26 TSNA with pre-existing T-cell responses. Importantly, because the dataset comprises largely neoantigen recognition by tumor-infiltrating lymphocytes, successful prediction implies the ability to identify not just neoantigens that are able to prime T-cells as in the literature[81, 82, 141], but – more stringently – neoantigens presented to T-cells by tumors.

[00494] We ranked mutations in order of probability of presentation using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the allele-specific neural network model described in Section VIII.B, and the pan-allele neural network model described in Section VIII.D. As capacities of antigen-specific immunotherapies are limited in the number of specificities targeted (e.g., current personalized vaccines encode ~10-20 mutations[6, 81, 82]), we compared predictive methods by counting the number of pre-existing T-cell responses in the top 5, 10, or 20-ranked mutations for each patient. These results are depicted in FIG. 25A. Specifically, FIG. 25A compares the proportion of somatic mutations recognized by T-cells (e.g., pre-existing T-cell responses) for the top 5, 10, and 20-ranked somatic mutations identified using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the

allele-specific neural network model, and the pan-allele neural network model for a test set comprising 12 different test samples, each test sample taken from a patient with at least one pre-existing T-cell response.

[00495] As expected, binding affinity prediction included only a minority of pre-existing T-cell responses among the prioritized mutations, for instance 9/26 (35%) among the top 20. In contrast, the majority (19/26, 73%) of pre-existing T-cell responses were ranked in the top 20 by both the allele-specific and the pan-allele NN models (FIG. 25A). These results confirm the pan-allele model's ability to identify human tumor CD8 T-cell epitopes with comparable accuracy (statistically insignificant) as the allele-specific model.

[00496] We then evaluated mutations at the level of minimal neoepitopes (i.e., which 8-11-mer overlapping the mutation was recognized), as may be useful to identify T-cells/TCRs for cell therapy. In other words, minimal neoepitopes were ranked in order of probability of presentation using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the allele-specific neural network model described in Section VIII.B, and the pan-allele neural network model described in Section VIII.D. As mentioned above, as antigen-specific immunotherapies are technically limited in the number of specificities targeted, predictive methods were compared by counting the number of pre-existing T-cell responses in the top 5, 10, or 20-ranked minimal neoepitopes for each patient with at least one pre-existing T-cell response. Positively-labeled epitopes were those confirmed to be immunogenic minimal epitopes via peptide-based (instead of, or in addition to, TMG-based assays), and negative examples were all epitopes not recognized in peptide-based assays and all mutation-spanning epitopes contained in non-recognized minigenes. The results are depicted in FIG. 25B.

[00497] Specifically, FIG. 25B compares the proportion of minimal neoepitopes recognized by T-cells (e.g., pre-existing T-cell responses) for the top 5, 10, and 20-ranked minimal neoepitopes identified using standard HLA binding affinity prediction with >2 TPM thresholds on gene expression as assayed by RNA-seq, the allele-specific neural network model, and the pan-allele neural network model for a test set comprising 12 different test samples, each test sample taken from a patient with at least one pre-existing T-cell response.

[00498] As shown in FIG. 25B, when evaluating mutations at the level of minimal epitopes, the pan-allele model continues to perform comparably to the allele-specific model.

XIV.A. Data

[00499] We obtained mutation calls, HLA types and T-cell recognition data from the supplementary information of Gros et al[84], Tran et al[140], Stronen et al[141] and Zacharakis et al. Patient-specific RNA-seq data were unavailable. Reasoning that tumor RNA expression is correlated across different patients with the same tumor type, RNA-seq data from tumor-type-matched patients from TCGA was substituted, which was used both in the neural network predictions and for RNA expression filtering before binding affinity prediction. The addition of tumor-type matched RNA-seq data improved predictive performance.

[00500] For the mutation-level analysis (FIG. 25A), the positive-labeled datapoints for Gros et al, Tran et al and Zacharakis et al were mutations recognized by patient T-cells in both the TMG assay or the minimal epitope peptide-pulsing assays. The negative-labeled datapoints were all other mutations tested in TMG assays. For Stronen et al, the positive labeled mutations were mutations spanned by at least one recognized peptide, and the negative datapoints were all mutations tested but not recognized in the tetramer assays. For the Gros, Tran and Zacharakis data, mutations were ranked either by summing probabilities of presentation or taking the minimum binding affinity across all mutation-spanning peptides, as the mutated-25mer TMG assay tests the T-cell recognition of all peptides spanning the mutation. For the Stronen data, mutations were ranked either by summing probabilities of presentation or taking the minimum binding affinity across all mutation-spanning peptides tested in the tetramer assays.

[00501] For the epitope-level analysis, the positive-labeled datapoints were all minimal epitopes recognized by patient T-cells in peptide-pulsing or tetramer assays, and the negative datapoints were all minimal epitopes not recognized by T-cells in peptide-pulsing or tetramer assays and all mutation-spanning peptides from tested TMGs that were not recognized by patient T-cells. In the case of Gros et al, Tran et al and Zacharakis et al minimal epitope peptides spanning mutations recognized in the TMG analysis that were not tested via peptide-pulsing assays were removed from the analysis, as the T-cell recognition status of these peptides was not determined experimentally.

XV. Example 11: Identification of Neoantigen-Reactive T-Cells in Cancer Patients

[00502] This example demonstrates that improved prediction can enable neoantigen identification from routine patient samples. To do so, archival FFPE tumor biopsies and 5-30ml of peripheral blood were analyzed from 9 patients with metastatic NSCLC undergoing

anti-PD(L)1 therapy (Supplementary Table 1: Patient demographics and treatment information for N=9 patients studied in FIGS. 26A-C. Key fields include tumor stage and subtype, anti-PD1 therapy received, and summary of NGS results.). Tumor whole exome sequencing, tumor transcriptome sequencing, and matched normal exome sequencing resulted in an average of 198 somatic mutations per patient (SNVs and short indel), of which an average of 118 were expressed (Methods, Supplementary Table 1). The full MS model was applied to prioritize 20 neoepitopes per patient for testing against pre-existing anti-tumor T-cell responses. To focus the analysis on likely CD8 responses, the prioritized peptides were synthesized as 8-11mer minimal epitopes (Methods), and then peripheral blood mononuclear cells (PBMCs) were cultured with the synthesized peptides in short *in vitro* stimulation (IVS) cultures to expand neoantigen-reactive T-cells (Supplementary Table 2). After two weeks the presence of antigen-specific T-cells was assessed using IFN-gamma ELISpot against the prioritized neoepitopes. In seven patients for whom sufficient PBMCs were available, separate experiments were also performed to fully or partially deconvolve the specific antigens recognized. The results are depicted in FIGS. 26A-C and 27A-30.

[00503] FIG. 26A depicts detection of T-cell responses to patient-specific neoantigen peptide pools for nine patients. For each patient, predicted neoantigens were combined into 2 pools of 10 peptides each according to model ranking and any sequence homologies (homologous peptides were separated into different pools). Then, for each patient, the *in vitro* expanded PBMCs for the patient were stimulated with the 2 patient-specific neoantigen peptide pools in IFN-gamma ELISpot. Data in FIG. 26A are presented as spot forming units (SFU) per 10^5 plated cells with background (corresponding DMSO negative controls) subtracted. Background measurements (DMSO negative controls) are shown in FIG. 30. Responses of single wells (patients 1-038-001, CU02, CU03 and 1-050-001) or replicates with mean and standard deviation (all other patients) against cognate peptide pools #1 and #2 are shown for patients 1-038-001, 1-050-001, 1-001-002, CU04, 1-024-001, 1-024-002 and CU05. For patients CU02 and CU03, cell numbers allowed testing against specific peptide pool #1 only. Samples with values >2-fold increase above background were considered positive and are designated with a star (responsive donors include patients 1-038-001, CU04, 1-024-001, 1-024-002, and CU02). Unresponsive donors include patients 1-050-001, 1-001-002, CU05, and CU03. FIG 15 C depicts photographs of ELISpot wells with *in vitro* expanded PBMCs from patient CU04, stimulated in IFN-gamma ELISpot with DMSO

negative control, PHA positive control, CU04-specific neoantigen peptide pool #1, CU04-specific peptide 1, CU04-specific peptide 6, and CU04-specific peptide 8.

[00504] FIGS. 27A-B depict results from control experiments with patient neoantigens in HLA-matched healthy donors. The results of these experiments verify that *in vitro* culture conditions expanded only pre-existing *in vivo* primed memory T-cells, rather than enabling *de novo* priming *in vitro*.

[00505] FIG. 28 depicts detection of T-cell responses to PHA positive control for each donor and each *in vitro* expansion depicted in FIG. 26A. For each donor and each *in vitro* expansion in FIG. 26A, the *in vitro* expanded patient PBMCs were stimulated with PHA for maximal T-cell activation. Data in FIG. 28 are presented as spot forming units (SFU) per 10^5 plated cells with background (corresponding DMSO negative controls) subtracted. Responses of single wells or biological replicates are shown for patients 1-038-001, 1-050-001, 1-001-002, CU04, 1-024-001, 1-024-002, CU05 and CU03. Testing with PHA was not conducted for patient CU02. Cells from patient CU02 were included into analyses, as a positive response against peptide pool #1 (FIG. 26A) indicated viable and functional T-cells. As shown in FIG. 26A, donors that were responsive to peptide pools include patients 1-038-001, CU04, 1-024-001, and 1-024-002. As also shown in FIG. 26A, donors that were unresponsive to peptide pools include patients 1-050-001, 1-001-002, CU05, and CU03.

[00506] FIG. 29A depicts detection of T-cell responses to each individual patient-specific neoantigen peptide in pool #2 for patient CU04. FIG. 29A also depicts detection of T-cell responses to PHA positive control for patient CU04. (This is positive control data is also shown in FIG. 28.) For patient CU04, the *in vitro* expanded PBMCs for the patient were stimulated in IFN-gamma ELISpot with patient-specific individual neoantigen peptides from pool #2 for patient CU04. The *in vitro* expanded PBMCs for the patient were also stimulated in IFN-gamma ELISpot with PHA as a positive control. Data are presented as spot forming units (SFU) per 10^5 plated cells with background (corresponding DMSO negative controls) subtracted.

[00507] FIG. 29B depicts detection of T-cell responses to individual patient-specific neoantigen peptides for each of three visits of patient CU04 and for each of two visits of patient 1-024-002, each visit occurring at a different time point. For both patients, the *in vitro* expanded PBMCs for the patient were stimulated in IFN-gamma ELISpot with patient-specific individual neoantigen peptides. For each patient, data for each visit are presented as cumulative (added) spot forming units (SFU) per 10^5 plated cells with background

(corresponding DMSO controls) subtracted. Data for patient CU04 are shown as background subtracted cumulative SFU from 3 visits. For patient CU04, background subtracted SFU are shown for the initial visit (T0) and subsequent visits 2 months (T0 + 2 months) and 14 months (T0 + 14 months) after the initial visit (T0). Data for patient 1-024-002 are shown as background subtracted cumulative SFU from 2 visits. For patient 1-024-002, background subtracted SFU are shown for the initial visit (T0) and a subsequent visit 1 month (T0 + 1 month) after the initial visit (T0). Samples with values >2-fold increase above background were considered positive and are designated with a star.

[00508] FIG. 29C depicts detection of T-cell responses to individual patient-specific neoantigen peptides and to patient-specific neoantigen peptide pools for each of two visits of patient CU04 and for each of two visits of patient 1-024-002, each visit occurring at a different time point. For both patients, the *in vitro* expanded PBMCs for the patient were stimulated in IFN-gamma ELISpot with patient-specific individual neoantigen peptides as well as with patient-specific neoantigen peptide pools. Specifically, for patient CU04, the *in vitro* expanded PBMCs for patient CU04 were stimulated in IFN-gamma ELISpot with CU04-specific individual neoantigen peptides 6 and 8 as well as with CU04-specific neoantigen peptide pools, and for patient 1-024-002, the *in vitro* expanded PBMCs for patient 1-024-002 were stimulated in IFN-gamma ELISpot with 1-024-002-specific individual neoantigen peptide 16 as well as with 1-024-002-specific neoantigen peptide pools. The data of FIG. 29C are presented as spot forming units (SFU) per 10^5 plated cells with background (corresponding DMSO controls) subtracted for each technical replicate with mean and range. Data for patient CU04 are shown as background subtracted SFU from 2 visits. For patient CU04, background subtracted SFU are shown for the initial visit (T0; technical triplicates) and a subsequent visit at 2 months (T0 + 2 months; technical triplicates) after the initial visit (T0). Data for patient 1-024-002 are shown as background subtracted SFU from 2 visits. For patient 1-024-002, background subtracted SFU are shown for the initial visit (T0; technical triplicates) and a subsequent visit 1 month (T0 + 1 month; technical duplicates, except for the sample stimulated with patient 1-024-002-specific neoantigen peptide pools) after the initial visit (T0).

[00509] FIG. 30 depicts detection of T-cell responses to the two patient-specific neoantigen peptide pools and to DMSO negative controls for the patients of FIG. 26A. For each patient, the *in vitro* expanded PBMCs for the patient were stimulated with the two patient-specific neoantigen peptide pools in IFN-gamma ELISpot. For each donor and each *in*

vitro expansion, the *in vitro* expanded patient PBMCs were also stimulated in IFN-gamma ELISpot with DMSO as a negative control. Data in FIG. 30 are presented as spot forming units (SFU) per 10^5 plated cells with background (corresponding DMSO negative controls) *included* for patient-specific neoantigen peptide pools and corresponding DMSO controls. Responses of single wells (1-038-001, CU02, CU03 and 1-050-001) or average with standard deviation of biological duplicates (all other samples) against cognate peptide pools #1 and #2 are shown for patients 1-038-001, 1-050-001, 1-001-002, CU04, 1-024-001, 1-024-002 and CU05. For patients CU02 and CU03, cell numbers allowed testing against specific peptide pool #1 only. Samples with values >2-fold increase above background were considered positive and are designated with a star (responsive donors include patients 1-038-001, CU04, 1-024-001, 1-024-002, and CU02). Unresponsive donors include patients 1-050-001, 1-001-002, CU05, and CU03.

[00510] As discussed briefly above with regard to FIGS. 27A-B, to verify that the *in vitro* culture conditions expanded only pre-existing *in vivo* primed memory T-cells, rather than enabling *de novo* priming *in vitro*, a series of control experiments were performed with neoantigens in HLA-matched healthy donors. The results of these experiments are depicted in FIGS. 27A-B and in Supplementary Table 4. The results of these experiments confirmed the absence of *de novo* priming and absence of a detectable neoantigen-specific T-cell response in healthy donors using IVS culture technique.

[00511] By contrast, pre-existing neoantigen-reactive T-cells were identified in the majority (5/9, 56%) of patients tested with patient-specific peptide pools (FIGS. 26A and 29-30) using IFN-gamma ELISpot. Of the 7 patients for whom cell numbers permitted complete or partial testing of individual neoantigen cognate peptides, 4 patients responded to at least one of the tested neoantigen peptides, and all of these patients had a corresponding pool response (FIG. 26B). The remaining 3 patients tested with individual neoantigens (patients 1-001-002, 1-050-001 and CU05) had no detectable responses against single peptides (data not shown), confirming the lack of response seen for these patients against neoantigen pools (FIG. 26A). Among the 4 responsive patients, samples from a single visit were available for 2 patients with a response (patients 1-024-001 and 1-038-001), while samples from multiple visits were available for the other 2 patients with a response (CU04 and 1-024-002). For the 2 patients with samples from multiple visits, the cumulative (added) spot forming units (SFU) from 3 visits (patient CU04) or 2 visits (patient 1-024-002) are shown in FIG. 26B and broken down by visit in FIG. 29B. Additional PBMC samples from the same visits were also

available for patients 1-024-002 and CU04, and repeat IVS culture and ELISpot confirmed responses to patient-specific neoantigens (FIG. 29C).

[00512] Overall, among patients for whom at least one T-cell recognized neoepitope was identified as shown by a response to a pool of 10 peptides in FIG. 26A, the number of recognized neoepitopes averaged at least 2 per patient (minimum of 10 epitopes identified in 5 patients, counting a recognized pool that could not be deconvolved as 1 recognized peptide). In addition to testing for IFN-gamma response by ELISpot, culture supernatants were also tested for granzyme B by ELISA and for TNF-alpha, IL-2 and IL-5 by MSD cytokine multiplex assay. Cells from 4 of the 5 patients with positive ELISpots secreted 3 or more analytes, including granzyme B (Supplementary Table 3), indicating polyfunctionality of neoantigen-specific T-cells. Importantly, because the combined prediction and IVS method did not rely on a limited set of available MHC multimers, responses were tested broadly across restricting HLA alleles. Furthermore, this approach directly identifies the minimal epitope, in contrast to tandem minigene screening, which identifies recognized mutations, and requires a separate deconvolution step to identify minimal epitopes. Overall, the neoantigen identification yield was comparable to previous best methods[96] testing TIL against all mutations with apheresis samples, while screening only 20 synthetic peptides with a routine 5-30mL of whole blood.

XV.A. Peptides

[00513] Custom-made, recombinant lyophilized peptides were purchased from JPT Peptide Technologies (Berlin, Germany) or Genscript (Piscataway, NJ, USA) and reconstituted at 10-50 mM in sterile DMSO (VWR International, Pittsburgh, PA, USA), aliquoted and stored at -80°C.

XV.B. Human Peripheral Blood Mononuclear Cells (PBMCs)

[00514] Cryopreserved HLA-typed PBMCs from healthy donors (confirmed HIV, HCV and HBV seronegative) were purchased from Precision for Medicine (Gladstone, NJ, USA) or Cellular Technology, Ltd. (Cleveland, OH, USA) and stored in liquid nitrogen until use. Fresh blood samples were purchased from Research Blood Components (Boston, MA, USA), leukopaks from AllCells (Boston, MA, USA) and PBMCs were isolated by Ficoll-Paque density gradient (GE Healthcare Bio, Marlborough, MA, USA) prior to cryopreservation. Patient PBMCs were processed at local clinical processing centers according to local clinical

standard operating procedures (SOPs) and IRB approved protocols. Approving IRBs were Quorum Review IRB, Comitato Etico Interaziendale A.O.U. San Luigi Gonzaga di Orbassano, and Comité Ético de la Investigación del Grupo Hospitalario Quirón en Barcelona.

[00515] Briefly, PBMCs were isolated through density gradient centrifugation, washed, counted, and cryopreserved in CryoStor CS10 (STEMCELL Technologies, Vancouver, BC, V6A 1B6, Canada) at 5 x 10^6 cells/ml. Cryopreserved cells were shipped in cryoports and transferred to storage in LN_2 upon arrival. Patient demographics are listed in Supplementary Table 1. Cryopreserved cells were thawed and washed twice in OpTmizer T-cell Expansion Basal Medium (Gibco, Gaithersburg, MD, USA) with Benzonase (EMD Millipore, Billerica, MA, USA) and once without Benzonase. Cell counts and viability were assessed using the Guava ViaCount reagents and module on the Guava easyCyte HT cytometer (EMD Millipore). Cells were subsequently re-suspended at concentrations and in media appropriate for proceeding assays (see next sections).

XV.C. *In vitro* stimulation (IVS) cultures

[00516] Pre-existing T-cells from healthy donor or patient samples were expanded in the presence of cognate peptides and IL-2 in a similar approach to that applied by Ott *et al.*[81] Briefly, thawed PBMCs were rested overnight and stimulated in the presence of peptide pools (10μM per peptide, 10 peptides per pool) in ImmunoCult™-XF T-cell Expansion Medium (STEMCELL Technologies) with 10 IU/ml rhIL-2 (R&D Systems Inc., Minneapolis, MN) for 14 days in 24-well tissue culture plates. Cells were seeded at 2 x 10^6 cells/well and fed every 2-3 days by replacing 2/3 of the culture media. One patient sample showed a deviation from the protocol and should be considered as a potential false negative: Patient CU03 did not yield sufficient numbers of cells post thawing and cells were seeded at 2 x 10^5 cells per peptide pool (10-fold fewer than per protocol).

XV.D. IFNγ Enzyme Linked Immunospot (ELISpot) assay

[00517] Detection of IFNγ-producing T-cells was performed by ELISpot assay[142]. Briefly, PBMCs (*ex vivo* or post *in vitro* expansion) were harvested, washed in serum free RPMI (VWR International) and cultured in the presence of controls or cognate peptides in OpTmizer T-cell Expansion Basal Medium (*ex vivo*) or in ImmunoCult™-XF T-cell

Expansion Medium (expanded cultures) in ELISpot Multiscreen plates (EMD Millipore) coated with anti-human IFNγ capture antibody (Mabtech, Cincinatti, OH, USA). Following 18h incubation in a 5% CO_2, 37°C, humidified incubator, cells were removed from the plate and membrane-bound IFNγ was detected using anti-human IFNγ detection antibody (Mabtech), Vectastain Avidin peroxidase complex (Vector Labs, Burlingame, CA, USA) and AEC Substrate (BD Biosciences, San Jose, CA, USA). ELISpot plates were allowed to dry, stored protected from light and sent to Zellnet Consulting, Inc., Fort Lee, NJ, USA) for standardized evaluation[143]. Data are presented as spot forming units (SFU) per plated number of cells.

XV.E. Granzyme B ELISA and MSD multiplex assay

[00518] Detection of secreted IL-2, IL-5 and TNF-alpha in ELISpot supernatants was performed using using a 3-plex assay MSD U-PLEX Biomarker assay (catalog number K15067L-2). Assays were performed according to the manufacturer's instructions. Analyte concentrations (pg/ml) were calculated using serial dilutions of known standards for each cytokine. For graphical data representation, values below the minimum range of the standard curve were represented equals zero. Detection of Granzyme B in ELISpot supernatants was performed using the Granzyme B DuoSet® ELISA (R & D Systems, Minneapolis, MN) according to the manufacturer's instructions. Briefly, ELISpot supernatants were diluted 1:4 in sample diluent and run alongside serial dilutions of Granzyme B standards to calculate concentrations (pg/ml). For graphical data representation, values below the minimum range of the standard curve were represented equals zero.

XV.F. Negative Control Experiments for IVS Assay – Neoantigens from Tumor Cell Lines Tested in Healthy Donors

[00519] FIG. 27A illustrates negative control experiments for IVS assay for neoantigens from tumor cell lines tested in healthy donors. Healthy donor PBMCs were stimulated in IVS culture with peptide pools containing positive control peptides (previous exposure to infectious diseases), HLA-matched neoantigens originating from tumor cell lines (unexposed), and peptides derived from pathogens for which the donors were seronegative. Expanded cells were subsequently analyzed by IFNγ ELISpot (10^5 cells/well) following stimulation with DMSO (negative controls, black circles), PHA and common infectious

diseases peptides (positive controls, red circles), neoantigens (unexposed, light blue circles), or HIV and HCV peptides (donors were confirmed to be seronegative, navy blue, A and B). Data are shown as spot forming units (SFU) per 10^5 seeded cells. Biological replicates with mean and SEM are shown. No responses were observed to neoantigens or to peptides deriving from pathogens to which the donors have not been exposed (seronegative).

XV.G. Negative Control Experiments for IVS Assay – Neoantigens from Patients Tested in Healthy Donors

[00520] FIG. 27A illustrates negative control experiments for IVS assay for neoantigens from patients tested for reactivity in healthy donors. Assessment of T-cell responses in healthy donors to HLA-matched neoantigen peptide pools. Left panel: Healthy donor PBMCs were stimulated with controls (DMSO, CEF and PHA) or HLA-matched patient-derived neoantigen peptides in *ex vivo* IFN-gamma ELISpot. Data are presented as spot forming units (SFU) per 2 x 10^5 plated cells for triplicate wells. Right panel: Healthy donor PBMCs post IVS culture, expanded in the presence of either neoantigen pool or CEF pool were stimulated in IFN-gamma ELISpot either with controls (DMSO, CEF and PHA) or HLA-matched patient-derived neoantigen peptide pool. Data are presented as SFU per 1 x 10^5 plated cells for triplicate wells. No responses to neoantigens in healthy donors are seen.

XV.H. Supplementary Table 2: Peptides Tested for T-Cell Recognition in NSCLC Patients

[00521] Details on neoantigen peptides tested for the N=9 patients studied in FIGS. 26A-C (Identification of Neoantigen-Reactive T-cells from NSCLC Patients). Key fields include source mutation, peptide sequence, and pool and individual peptide responses observed. The "most_probable_restriction" column indicates which allele the model predicted was most likely to present each peptide. The ranks of these peptides among all mutated peptides for each patient as computed with binding affinity prediction (Methods) are also included.

[00522] There were four peptides highly ranked by the full MS model and recognized by CD8 T-cells that had low predicted binding affinities or were ranked low by binding affinity prediction.

[00523] For three of these peptides, this is caused by differences in HLA coverage between the model and MHCflurry 1.2.0. Peptide YEHEDVKEA is predicted to be presented by HLA-B*49:01, which is not covered by MHCflurry 1.2.0. Similarly, peptides SSAAAPFPL and FVSTSDIKSM are predicted to be presented by HLA-C*03:04,

which is also not covered by MHCflurry 1.2.0. The online NetMHCpan 4.0 (BA) predictor, a pan-specific binding affinity predictor that in principle covers all alleles, ranks SSAAAPFPL as a strong binder to HLA-C*03:04 (23.2nM, ranked 2nd for patient 1-024-002), predicts weak binding of FVSTSDIKSM to HLA-C*03:04 (943.4nM, ranked 39th for patient 1-024-002) and weak binding of YEHEDVKEA to HLA-B*49:01 (3387.8nM), but stronger binding to HLA-B*41:01 (208.9nM, ranked 11th for patient 1-038-001), which is also present in this patient but is not covered by the model. Thus, of these three peptides, FVSTSDIKSM would have been missed by binding affinity prediction, SSAAAPFPL would have been captured, and the HLA restriction of YEHEDVKEA is uncertain.

[00524] The remaining five peptides for which a peptide-specific T-cell response was deconvolved came from patients where the most probable presenting allele as determined by the model was also covered by MHCflurry 1.2.0. Of these five peptides, 4/5 had predicted binding affinities stronger than the standard 500nM threshold and ranked in the top 20, though with somewhat lower ranks than the ranks from the model (peptides DENITTIQF, QDVSVQVER, EVADAATLTM, DTVEYPYTSF were ranked 0, 4, 5, 7 by the model respectively vs 2, 14, 7, and 9 by MHCflurry). Peptide GTKKDVDVLK was recognized by CD8 T-cells and ranked 1 by the model, but had rank 70 and predicted binding affinity 2169 nM by MHCflurry.

[00525] Overall, 6/8 of the individually-recognized peptides that were ranked highly by the full MS model also ranked highly using binding affinity prediction and had predicted binding affinity <500nM, while 2/8 of the individually-recognized peptides would have been missed if binding affinity prediction had been used instead of the full MS model.

XV.I. Supplementary Table 3: MSD Cytokine Multiplex and ELISA Assays on ELISpot Supernatants from NSCLC Neoantigen Peptides

[00526] Analytes detected in supernatants from positive ELISpot (IFNgamma) wells are shown for granzyme B (ELISA), TNFalpha, IL-2 and IL-5 (MSD). Values are shown as average pg/ml from technical replicates. Positive values are shown in italics. Granzyme B ELISA: Values ≥1.5-fold over DMSO background were considered positive. U-Plex MSD assay: Values ≥1.5-fold over DMSO background were considered positive.

XV.J. Supplementary Table 4: Neoantigen and Infectious Disease Epitopes in IVS Control Experiments

[00527] Details on tumor cell line neoantigen and viral peptides tested in IVS control experiments shown in FIGS. 27A-B. Key fields include source cell line or virus, peptide sequence, and predicted presenting HLA allele.

XV.K. Data

[00528] The MS peptide dataset used to train and test the prediction model (FIGS. 25A-B) is available at the MassIVE Archive (massive.ucsd.edu), accession number MSV000082648. Neoantigen peptides tested by ELISpot (FIGS. 26A-C and 27A-B) are included with the manuscript (Supplementary Tables 2 and 4).

XVI. Methods of Examples 8-11

XVI.A. Mass Spectrometry

XVI.A.1. Specimens

[00529] Archived frozen tissue specimens for mass spectrometry analysis were obtained from commercial sources, including BioServe (Beltsville, MD), ProteoGenex (Culver City, CA), iSpecimen (Lexington, MA), and Indivumed (Hamburg, Germany). A subset of specimens was also collected prospectively from patients at Hopital Marie Lannelongue (Le Plessis-Robinson, France) under a research protocol approved by the Comité de Protection des Personnes, Ile-de-France VII.

XVI.A.2. HLA Immunoprecipitation

[00530] Isolation of HLA-peptide molecules was performed using established immunoprecipitation (IP) methods after lysis and solubilization of the tissue sample[87,124–126]. Fresh frozen tissue was pulverized (CryoPrep; Covaris, Woburn, MA), lysis buffer (1% CHAPS, 20mM Tris-HCl, 150mM NaCl, protease and phosphatase inhibitors, pH=8) was added to solubilize the tissue and the resultant solution was centrifuged at 4C for 2 hrs to pellet debris. The clarified lysate is used for the HLA specific IP. Immunoprecipitation was performed as previously described using the antibody W6/32[127]. The lysate is added to the antibody beads and rotated at 4C overnight for the immunoprecipitation. After

immunoprecipitation, the beads were removed from the lysate. The IP beads were washed to remove non-specific binding and the HLA/peptide complex was eluted from the beads with 2N acetic acid. The protein components were removed from the peptides using a molecular weight spin column. The resultant peptides were taken to dryness by SpeedVac evaporation and stored at -20C prior to MS analysis.

XVI.A.3. Peptide Sequencing

[00531] Dried peptides were reconstituted in HPLC buffer A and loaded onto a C-18 microcapillary HPLC column for gradient elution in to the mass spectrometer. A gradient of 0-40%B (solvent A – 0.1% formic acid, solvent B- 0.1% formic acid in 80% acetonitrile) in 180 minutes was used to elute the peptides into the Fusion Lumos mass spectrometer (Thermo). MS1 spectra of peptide mass/charge (m/z) were collected in the Orbitrap detector with 120,000 resolution followed by 20 MS2 low resolution scans collected in the either the Orbitrap or ion trap detector after HCD fragmentation of the selected ion. Selection of MS2 ions was performed using data dependent acquisition mode and dynamic exclusion of 30 seconds after MS2 selection of an ion. Automatic gain control (AGC) for MS1 scans was set to 4x105 and for MS2 scans was set to 1x104. For sequencing HLA peptides, +1, +2 and +3 charge states can be selected for MS2 fragmentation.

[00532] MS2 spectra from each analysis were searched against a protein database using Comet[128,129] and the peptide identification were scored using Percolator[130–132].

XVI.B. Machine Learning

XVI.B.1. Data Encoding

[00533] For each sample, the training data points were all 8-11mer (inclusive) peptides from the reference proteome that mapped to exactly one gene expressed in the sample. The overall training dataset was formed by concatenating the training datasets from each training sample. Lengths 8-11 were chosen as this length range captures ~95% of all HLA class I presented peptides; however, adding lengths 12-15 to the model could be accomplished using the same methodology, at the cost of a modest increase in computational demands. Peptides and flanking sequence were vectorized using a one-hot encoding scheme. Peptides of multiple lengths (8-11) were represented as fixed-length vectors by augmenting the amino acid alphabet with a pad character and padding all peptides to the maximum length of 11.

RNA abundance of the source protein of the training peptides was represented as the logarithm of the isoform-level transcripts per million (TPM) estimate obtained from RSEM[133]. For each peptide, the per-peptide TPM was computed as the sum of the per-isoform TPM estimates for each of the isoforms that contain the peptide. Peptides from genes expressed at 0 TPM were excluded from the training data, and at test time, peptides from non-expressed genes are assigned a probability of presentation of 0. Lastly, each peptide was assigned to an Ensembl protein family ID, and each unique Ensembl protein family ID corresponded to a per-gene presentation propensity intercept (see next section).

XVII.B.2. Specification of the Model Architecture

[00534] The full presentation model has the following functional form:

(Equation 1) $Pr(peptide\ i\ presented) = \sum_{k=1}^{m} a_k^i \cdot Pr(peptide\ i\ presented\ by\ allele\ a)$,

where k indexes HLA alleles in the dataset, which run from 1 to m, and a_k^i is an indicator variable whose value is 1 if allele k is present in the sample from which peptide i is derived and 0 otherwise. Note that for a given peptide i, all but at most 6 of the a_k^i (the 6 corresponding to the HLA type of the sample of origin of peptide i) will be zero. The sum of probabilities is clipped at $1 - \epsilon$, with $\epsilon = 10^{-6}$ for instance.

[00535] The per-allele probabilities of presentation are modeled as below:

$$\Pr(\text{peptide } i \text{ presented by allele } a) = \text{sigmoid}\left\{NN_a(\text{peptide}_i) + NN_{\text{flanking}}(\text{flanking}_i) + NN_{\text{RNA}}(\log(\text{TPM}_i)) + \alpha_{\text{sample}(i)} + \beta_{\text{protein(i)}}\right\},$$

where the variables have the following meanings: sigmoid is the sigmoid (aka expit) function, peptide_i is the onehot-encoded middle-padded amino acid sequence of peptide i, NN_a is a neural network with linear last-layer activation modeling the contribution of the peptide sequence to the probability of presentation, flanking_i is the onehot-encoded flanking sequence of peptide i in its source protein, NN_{flanking} is a neural network with linear last-layer activation modeling the contribution of the flanking sequence to the probability of presentation, TPM_i is the expression of the source mRNAs of peptide i in TPM units, $\text{sample}(i)$ is the sample (i.e., patient) of origin of peptide i, $\alpha_{\text{sample}(i)}$ is a per-sample intercept, protein(i) is the source protein of peptide i, and $\beta_{\text{protein(i)}}$ is a per-protein intercept (aka the per-gene propensity of presentation).

[00536] For the models described in the results section, the component neural networks have the following architectures:

- Each of the NN_a is one output node of a one-hidden-layer multi-layer-perceptron (MLP) with input dimension 231 (11 residues x 21 possible characters per residue, including the pad character), width 256, rectified linear unit (ReLU) activations in the hidden layer, linear activation in the output layer, and one output node per HLA allele a in the training dataset.
- NN_{flanking} is a one- hidden-layer MLP with input dimension 210 (5 residues of N-terminal flanking sequence + 5 residues of C-terminal flanking sequence x 21 possible characters per residue, including the pad character), width 32, rectified linear unit (ReLU) activations in the hidden layer and linear activation in the output layer.
- NN_{RNA} is a one- hidden-layer MLP with input dimension 1, width 16, rectified linear unit (ReLU) activations in the hidden layer and linear activation in the output layer.

[00537] Note that some components of the model (e.g., NN_a) depend on a particular HLA allele, but many components (NN_{flanking}, NN_{RNA}, $\alpha_{\text{sample}(i)}$, $\beta_{\text{protein(i)}}$) do not. The former is referred to as "allele-interacting" and the latter as "allele-noninteracting". Features to model as allele-interacting or noninteracting were chosen on the basis of biological prior knowledge: the HLA allele sees the peptide, so the peptide sequence should be modeled as allele-interacting, but no information about the source protein, RNA expression or flanking sequence is conveyed to the HLA molecule (as the peptide has been separated from its source protein by the time it encounters the HLA in the endoplasmic reticulum), so these features should be modeled as allele-noninteracting. The model was implemented in Keras v2.0.4[134] and Theano v0.9.0[135].

[00538] The peptide MS model used the same deconvolution procedure as the full MS model (Equation 1), but the per-allele probabilities of presentation were generated using reduced per-allele models that consider only peptide sequence and HLA allele:

$$\Pr(\text{peptide } i \text{ presented by allele } a) = \text{sigmoid}\{NN_a(\text{peptide}_i)\}.$$

[00539] The peptide MS model uses the same features as binding affinity prediction, but the weights of the model are trained on a different data type (i.e., mass spectrometry data vs HLA-peptide binding affinity data). Therefore, comparing the predictive performance of the peptide MS model to the full MS model reveals the contribution of non-peptide features (i.e.,

RNA abundance, flanking sequence, gene ID) to the overall predictive performance, and comparing the predictive performance of the peptide MS model to the binding affinity models reveals the importance of improved modeling of the peptide sequence to the overall predictive performance.

XVI.B.3. Train/ Validate/ Test Splits

We ensured that no peptides appeared in more than one of the training / validation / testing sets using the following procedure: first by removing all peptides from the reference proteome that appear in more than one protein, then by partitioning the proteome into blocks of 10 adjacent peptides. Each block was assigned uniquely to the training, validation or testing sets. In this way, no peptide appears in more than one of the training, validation on testing sets. The validation set was used only for early stopping.

XVI.B.4. Model Training

[00540] For model training, all peptides were modeled as independent where the per-peptides loss is the negative Bernoulli log-likelihood loss function (aka log loss). Formally, the contribution of peptide i to the overall loss is

$$\text{Loss}(i) = -\log\Big(\text{Bernoulli}\big(y_i \mid \Pr(\text{peptide } i \text{ presented})\big)\Big),$$

where y_i is the label of peptide i; i.e., $y_i = 1$ if peptide i is presented and 0 otherwise, and $\text{Bernoulli}(y \mid p)$ detnoes the Bernoulli likelihood of parameter $p \in [0, 1]$ given i.i.d. binary observation vector y. The model was trained by minimizing the loss function.

[00541] In order to reduce training time, the class balance was adjusted by removing 90% of the negative-labeled training data at random, yielding an overall training set class balance of one presented peptide per ~2000 non-presented peptides. Model weights were initialized using the Glorot uniform procedure61 and trained using the ADAM62 stochastic optimizer with standard parameters on Nvidia Maxwell TITAN X GPUs. A validation set consisting of 10% of the total data was used for early stopping. The model was evaluated on the validation set every quarter-epoch and model training was stopped after the first quarter-epoch where the validation loss (i.e., the negative Bernoulli log-likelihood on the validation set) failed to decrease.

[00542] The full presentation model was an ensemble of 10 model replicates, with each replicate trained independently on a shuffled copy of the same training data with a different

random initialization of the model weights for every model within the ensemble. At test time, predictions were generated by taking the mean of the probabilities output by the model replicates.

XVI.B.5. Motif Logos

[00543] Motif logos were generated using the weblogolib Python API v3.5.0[138]. To generate binding affinity logos, the mhc_ligand_full.csv file was downloaded from the Immune Epitope Database (IEDB[88]) in July, 2017 and peptides meeting the following criteria were retained: measurement in nanomolar (nM) units, reference date after 2000, object type equal to "linear peptide" and all residues in the peptide drawn from the canonical 20-letter amino acid alphabet. Logos were generated using the subset of the filtered peptides with measured binding affinity below the conventional binding threshold of 500nM. For alleles pair with too few binders in IEDB, logos were not generated. To generate logos representing the learned presentation model, model predictions for 2,000,000 random peptides were predicted for each allele and each peptide length. For each allele and each length, the logos were generated using the peptides ranked in the top 1% (i.e., the top 20,000) by the learned presentation model. Importantly, this binding affinity data from IEDB was not used in model training or testing, but rather used only for the comparison of motifs learned.

XVI.B.6. Binding Affinity Prediction

[00544] We predicted peptide-MHC binding affinities using the binding affinity-only predictor from MHCflurry v1.2.0[139], an open-source, GPU-compatible HLA class I binding affinity predictor with performance comparable to the NetMHC family of models. To combine binding affinity predictions for a single peptide across multiple HLA alleles, the minimum binding affinity was selected. To combine binding affinities across multiple peptides (i.e., in order to rank mutations spanned by multiple mutated peptides as in FIGS. 25A-B), the minimum binding affinity across the peptides was selected. For RNA expression thresholding on the T-cell dataset, tumor-type matched RNA-seq data from TCGA to threshold at TPM>1 was used. All of the original T-cell datasets were filtered on TPM>0 in the original publications, so the TCGA RNA-seq data to filter on TPM>0 was not used.

XVI.B.7. Presentation Prediction

[00545] To combine probabilities of presentation for a single peptide across multiple HLA alleles, the sum of the probabilities was identified, as in Equation 1. To combine probabilities of presentation across multiple peptides (i.e., in order to rank mutations spanned by multiple peptides as in FIGS. 25A-B), the sum of the probabilities of presentation was identified. Probabilistically, if presentation of the peptides is viewed as i.i.d. Bernoulli random variables, the sum of the probabilities corresponds to the expected number of presented mutated peptides:

$$\text{E[\# presented neoantigens spanning mutation } i] = \sum_{j=1}^{n_i} \text{Pr[epitope } j \text{ presented]},$$

where Pr[epitope j presented] is obtained by applying the trained presentation model to epitope j, and n_i denotes the number of mutated epitopes spanning mutation i. For example, for an SNV i distant from the termini of its source gene, there are 8 spanning 8-mers, 9-spanning 9-mers, 10 spanning 10-mers and 11 spanning 11-mers, for a total of $n_i = 38$ spanning mutated epitopes.

XVI.C. Next Generation Sequencing

XVI.C.1. Specimens

[00546] For transcriptome analysis of the frozen resected tumors, RNA was obtained from same tissue specimens (tumor or adjacent normal) as used for MS analyses. For neoantigen exome and transcriptome analysis in patients on anti-PD1 therapy, DNA and RNA was obtained from archival FFPE tumor biopsies. Adjacent normal, matched blood or PBMCs were used to obtain normal DNA for normal exome and HLA typing.

XVI.C.2. Nucleic Acid Extraction and Library Construction

[00547] Normal/germline DNA derived from blood were isolated using Qiagen DNeasy columns (Hilden, Germany) following manufacturer recommended procedures. DNA and RNA from tissue specimens were isolated using Qiagen Allprep DNA/RNA isolation kits following manufacturer recommended procedures. The DNA and RNA were quantitated by Picogreen and Ribogreen Fluorescence (Molecular Probes), respectively specimens with >50ng yield were advanced to library construction. DNA sequencing libraries were generated

by acoustic shearing (Covaris, Woburn, MA) followed by DNA Ultra II (NEB, Beverly, MA) library preparation kit following the manufacturers recommended protocols. Tumor RNA sequencing libraries were generated by heat fragmentation and library construction with RNA Ultra II (NEB). The resulting libraries were quantitated by Picogreen (Molecular Probes).

XVI.C.3. Whole Exome Capture

[00548] Exon enrichment for both DNA and RNA sequencing libraries was performed using xGEN Whole Exome Panel (Integrated DNA Technologies). One to 1.5 µg of normal DNA or tumor DNA or RNA-derived libraries were used as input and allowed to hybridize for greater than 12 hours followed by streptavidin purification. The captured libraries were minimally amplified by PCR and quantitated by NEBNext Library Quant Kit (NEB). Captured libraries were pooled at equimolar concentrations and clustered using the c-bot (Illumina) and sequenced at 75 base paired-end on a HiSeq4000 (Illumina) to a target unique average coverage of >500x tumor exome, >100x normal exome, and >100M reads tumor transcriptome.

XVI.C.4. Analysis

[00549] Exome reads (FFPE tumor and matched normals) were aligned to the reference human genome (hg38) using BWA-MEM[144] (v. 0.7.13-r1126). RNA-seq reads (FFPE and frozen tumor tissue samples) were aligned to the genome and GENCODE transcripts (v. 25) using STAR (v. 2.5.1b). RNA expression was quantified using RSEM[133] (v. 1.2.31) with the same reference transcripts. Picard (v. 2.7.1) was used to mark duplicate alignments and calculate alignment metrics. For FFPE tumor samples following base quality score recalibration with GATK[145] (v. 3.5-0), substitution and short indel variants were determined using paired tumor-normal exomes with FreeBayes[146] (1.0.2). Filters included allele frequency >4%; median base quality >25, minimum mapping quality of supporting reads 30, and alternate read count in normal <=2 with sufficient coverage obtained. Variants must also be detected on both strands. Somatic variants occurring in repetitive regions were excluded. Translation and annotation were performed with snpEff[147] (v. 4.2) using RefSeq transcripts. Non-synonymous, non-stop variants verified in tumor RNA alignments were advanced to neoantigen prediction. Optitype[148] 1.3.1 was used to generate HLA types.

XVI.C.5. FIGS. 27A-B: Tumor Cell Lines and Matched Normals for IVS Control Experiments

[00550] Tumor cell lines H128, H122, H2009, H2126, Colo829 and their normal donor matched control cell lines BL128, BL2122, BL2009, BL2126 and Colo829BL were all purchased from ATCC (Manassas, VA) were grown to 10^{83}-10^{84} cells per seller's instructions then snap frozen for nucleic acid extraction and sequencing. NGS processing was performed generally as described above, except that MuTect[149] (3.1-0) was used for substitution mutation detection only. Peptides used in the IVS control assays are listed in Supplementary Table 4.

XVI.D. Class II Model Proof-of-Concept

[00551] To demonstrate the ability of the pan-allele neural network (NN) model to predict presentation by MHC class II molecules, an experiment was conducted using human B cell lymphomas samples (n=39). Each of the 39 samples comprised HLA-DR molecules, more specifically, HLA-DRB1 molecules, HLA-DRB3 molecules, HLA-DRB4 molecules, and/or HLA-DRB5 molecules. Four of the samples were set aside as a testing set and the other 35 samples were used for training and validation. The training set consisted of 20,136 presented peptides of 9-20 amino acids (AA) in length, inclusive, with modes of 13 and 14 amino acids long. The validation set and the test set consisted of 2,279 and 301 presented peptides, respectively.

[00552] The MHC class II pan-allele NN model architecture was identical to the MHC class I pan-allele NN model architecture, with 3 exceptions: (1) the class II model accepted up to 4 unique HLA-DRB alleles per sample (instead of 6 alleles of HLA-A, HLA-B, HLA-C), (2) the class II model was trained on longer peptide sequences, 9-20mers instead of 8-11mers, and (3) the per-allele model fit a distinct sub-network model for each allele whereas the pan-allele model shared knowledge between alleles by using a shared dense network for all alleles. The performance of the pan-allele model was compared against the allele-specific NN model. Both models were trained on the same peptides. The only difference to the model input between the two NN models was that the pan-allele model used a 34 length AA sequence to describe the HLA types whereas the allele-specific model used the standard HLA nomenclature (e.g., HLA-DRB1*01:01).

[00553] FIGS. 31A-D display the precision-recall curves for each of the test samples for the pan-allele and the allele-specific models. Specifically, FIG. 31A depicts the precision-

recall curves for each of the test sample 0 for the pan-allele and the allele-specific models. FIG. 31B depicts the precision-recall curves for each of the test sample 1 for the pan-allele and the allele-specific models. FIG. 31C depicts the precision-recall curves for each of the test sample 2 for the pan-allele and the allele-specific models. FIG. 31D depicts the precision-recall curves for each of the test sample 4 for the pan-allele and the allele-specific models. As shown in FIGS. 31A-D, both NN models achieve comparable (statistically insignificant) positive predictive value scores, and likewise for area under the receiver operating characteristic curve (ROC AUC) (see also Tables 3 and 4 below). This demonstrates the pan-allele model's ability to match the performance of an allele-specific model in the task of MHC class II peptide presentation prediction.

	Positive Predictive Value at 40% Recall				
	Test Sample 0	Test Sample 1	Test Sample 2	Test Sample 3	Mean
Allele-Specific	6.4%	6.5%	1.4%	6.8%	5.3%
Pan-Allele	3.8%	10.9%	3.9%	3.2%	5.4%

Table 3

	Area Under the ROC Curve				
	Test Sample 0	Test Sample 1	Test Sample 2	Test Sample 3	Mean
Allele-Specific	0.99	0.98	0.96	0.97	0.98
Pan-Allele	0.99	0.98	0.98	0.98	0.98

Table 4

XVII. Example 12: Sequencing TCRs of Neoantigen-Specific Memory T-Cells from Peripheral Blood of a NSCLC Patient

[00554] FIG. 32 depicts a method for sequencing TCRs of neoantigen-specific memory T-cells from the peripheral blood of a NSCLC patient. Peripheral blood mononuclear cells (PBMCs) from NSCLC patient CU04 (described above with regard to FIGS. 26A-30) were collected after ELISpot incubation. Specifically, as discussed above, the *in vitro* expanded PBMCs from 2 visits of patient CU04 were stimulated in IFN-gamma ELISpot with the CU04-specific individual neoantigen peptides (FIG. 29C), with the CU04-specific neoantigen

peptide pool (FIG. 29C), and with DMSO negative control (FIG. 30). Following incubation and prior to addition of detection antibody, the PBMCs were transferred to a new culture plate and maintained in an incubator during completion of the ELISpot assay. Positive (responsive) wells were identified based on ELISpot results. As shown in FIG. 32, the positive wells identified include the wells stimulated with CU04-specific individual neoantigen peptide 8 and the wells simulated with the CU04-specific neoantigen peptide pool. Cells from these positive wells and negative control (DMSO) wells were combined and stained for CD137 with magnetically-labelled antibodies for enrichment using Miltenyi magnetic isolation columns.

[00555] CD137-enriched and -depleted T-cell fractions isolated and expanded as described above were sequenced using 10x Genomics single cell resolution paired immune TCR profiling approach. Specifically, live T cells were partitioned into single cell emulsions for subsequent single cell cDNA generation and full-length TCR profiling (5' UTR through constant region – ensuring alpha and beta pairing). One approach utilizes a molecularly barcoded template switching oligo at the 5'end of the transcript, a second approach utilizes a molecularly barcoded constant region oligo at the 3' end, and a third approach couples an RNA polymerase promoter to either the 5' or 3' end of a TCR. All of these approaches enable the identification and deconvolution of alpha and beta TCR pairs at the single-cell level. The resulting barcoded cDNA transcripts underwent an optimized enzymatic and library construction workflow to reduce bias and ensure accurate representation of clonotypes within the pool of cells. Libraries were sequenced on Illumina's MiSeq or HiSeq4000 instruments (paired-end 150 cycles) for a target sequencing depth of about five to fifty thousand reads per cell. The resulting TCR nucleic acid sequences are depicted in Supplementary Table 5. The presence of the TCRa and TCRb chains described in Supplementary Table 5 were confirmed by an orthogonal anchor-PCR based TCR sequencing approach (Archer). This particular approach has the advantage of using limited cell numbers as input and fewer enzymatic manipulations when compared to the 10x Genomics based TCR sequencing.

[00556] Sequencing outputs were analyzed using the 10x software and custom bioinformatics pipelines to identify T-cell receptor (TCR) alpha and beta chain pairs as also shown in Supplementary Table 5. Supplementary Table 5 further lists the alpha and beta variable (V), joining (J), constant (C), and beta diversity (D) regions, and CDR3 amino acid sequence of the most prevalent TCR clonotypes. Clonotypes were defined as alpha, beta

chain pairs of unique CDR3 amino acid sequences. Clonotypes were filtered for single alpha and single beta chain pairs present at frequency above 2 cells to yield the final list of clonotypes per target peptide in patient CU04 (Supplementary Table 5).

[00557] In summary, using the method described above with regard to FIG. 32, memory CD8+ T-cells from the peripheral blood of patient CU04, that are neoantigen-specific to patient CU04's tumor neoantigens identified as discussed above with regard to Example 11 in Section XV., were identified. The TCRs of these identified neoantigen-specific T-cells were sequenced. And furthermore, sequenced TCRs that are neoantigen-specific to patient CU04's tumor neoantigens as identified by the above presentation models, were identified.

XVIII. Example 13: Use of Neoantigen-Specific Memory T-Cells for T-Cell Therapy

[00558] After T-cells and/or TCRs that are neoantigen-specific to neoantigens presented by a patient's tumor are identified, these identified neoantigen-specific T-cells and/or TCRs can be used for T-cell therapy in the patient. Specifically, these identified neoantigen-specific T-cells and/or TCRs can be used to produce a therapeutic quantity of neoantigen-specific T-cells for infusion into a patient during T-cell therapy. Two methods for producing a therapeutic quantity of neoantigen specific T-cells for use in T-cell therapy in a patient are discussed herein in Sections XVIII.A. and XVIII.B. The first method comprises expanding the identified neoantigen-specific T-cells from a patient sample (Section XVIII.A.). The second method comprises sequencing the TCRs of the identified neoantigen-specific T-cells and cloning the sequenced TCRs into new T-cells (Section XVIII.B.). Alternative methods for producing neoantigen specific T-cells for use in T-cell therapy that are not explicitly mentioned herein can also be used to produce a therapeutic quantity of neoantigen specific T-cells for use in T-cell therapy. Once the neoantigen-specific T-cells are obtained via one or more of these methods, these neoantigen-specific T-cells may be infused into the patient for T-cell therapy.

XVIII.A. Identification and Expansion of Neoantigen-Specific Memory T-Cells from a Patient Sample for T-Cell Therapy

[00559] A first method for producing a therapeutic quantity of neoantigen specific T-cells for use in T-cell therapy in a patient comprises expanding identified neoantigen-specific T-cells from a patient sample.

[00560] Specifically, to expand neoantigen-specific T-cells to a therapeutic quantity for use in T-cell therapy in a patient, a set of neoantigen peptides that are most likely to be presented by a patient's cancer cells are identified using the presentation models as described above. Additionally, a patient sample containing T-cells is obtained from the patient. The patient sample may comprise the patient's peripheral blood, tumor-infiltrating lymphocytes (TIL), or lymph node cells.

[00561] In embodiments in which the patient sample comprises the patient's peripheral blood, the following methods may be used to expand neoantigen-specific T-cells to a therapeutic quantity. In one embodiment, priming may be performed. In another embodiment, already-activated T-cells may be identified using one or more of the methods described above. In another embodiment, both priming and identification of already-activated T-cells may be performed. The advantage to both priming and identifying already-activated T-cells is to maximize the number of specificities represented. The disadvantage both priming and identifying already-activated T-cells is that this approach is difficult and time-consuming. In another embodiment, neoantigen-specific cells that are not necessarily activated may be isolated. In such embodiments, antigen-specific or non-specific expansion of these neoantigen-specific cells may also be performed. Following collection of these primed T-cells, the primed T-cells can be subjected to rapid expansion protocol. For example, in some embodiments, the primed T-cells can be subjected to the Rosenberg rapid expansion protocol (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2978753/, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2305721/)[153, 154].

[00562] In embodiments in which the patient sample comprises the patient's TIL, the following methods may be used to expand neoantigen-specific T-cells to a therapeutic quantity. In one embodiment, neoantigen-specific TIL can be tetramer/multimer sorted *ex vivo*, and then the sorted TIL can be subjected to a rapid expansion protocol as described above. In another embodiment, neoantigen-nonspecific expansion of the TIL may be performed, then neoantigen-specific TIL may be tetramer sorted, and then the sorted TIL can be subjected to a rapid expansion protocol as described above. In another embodiment, antigen-specific culturing may be performed prior to subjecting the TIL to the rapid expansion protocol. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607110/, https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.201545849)[155, 156].

[00563] In some embodiments, the Rosenberg rapid expansion protocol may be modified. For example, anti-PD1 and/or anti-41BB may be added to the TIL culture to simulate more rapid expansion. (https://jitc.biomedcentral.com/articles/10.1186/s40425-016-0164-7)[157].

XVIII.B. Identification of Neoantigen-Specific T Cells, Sequencing TCRs of Identified Neoantigen-Specific T Cells, and Cloning of Sequenced TCRs into new T-Cells

[00564] A second method for producing a therapeutic quantity of neoantigen specific T-cells for use in T-cell therapy in a patient comprises identifying neoantigen-specific T-cells from a patient sample, sequencing the TCRs of the identified neoantigen-specific T-cells, and cloning the sequenced TCRs into new T-cells.

[00565] First, neoantigen-specific T-cells are identified from a patient sample, and the TCRs of the identified neoantigen-specific T-cells are sequenced. The patient sample from which T cells can be isolated may comprise one or more of blood, lymph nodes, or tumors. More specifically, the patient sample from which T cells can be isolated may comprise one or more of peripheral blood mononuclear cells (PBMCs), tumor-infiltrating cells (TILs), dissociated tumor cells (DTCs), *in vitro* primed T cells, and/or cells isolated from lymph nodes. These cells may be fresh and/or frozen. The PBMCs and the *in vitro* primed T cells may be obtained from cancer patients and/or healthy subjects.

[00566] After the patient sample is obtained, the sample may be expanded and/or primed. Various methods may be implemented to expand and prime the patient sample. In one embodiment, fresh and/or frozen PBMCs may be simulated in the presence of peptides or tandem mini-genes. In another embodiment, fresh and/or frozen isolated T-cells may be simulated and primed with antigen-presenting cells (APCs) in the presence of peptides or tandem mini-genes. Examples of APCs include B-cells, monocytes, dendritic cells, macrophages or artificial antigen presenting cells (such as cells or beads presenting relevant HLA and co-stimulatory molecules, reviewed in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929753). In another embodiment, PBMCs, TILs, and/or isolated T-cells may be stimulated in the presence of cytokines (e.g., IL-2, IL-7, and/or IL-15). In another embodiment, TILs and/or isolated T-cells can be stimulated in the presence of maximal stimulus, cytokine(s), and/or feeder cells. In such embodiments, T cells can be isolated by activation markers and/or multimers (e.g., tetramers). In another embodiment, TILs and/or isolated T cells can be stimulated with stimulatory and/or co-stimulatory markers (e.g., CD3 antibodies, CD28 antibodies, and/or beads (e.g., DynaBeads). In

another embodiment, DTCs can be expanded using a rapid expansion protocol on feeder cells at high dose of IL-2 in rich media.

[00567] Then, neoantigen-specific T cells are identified and isolated. In some embodiments, T cells are isolated from a patient sample *ex vivo* without prior expansion. In one embodiment, the methods described above with regard to Section XVII. may be used to identify neoantigen-specific T cells from a patient sample. In an alternative embodiment, isolation is carried out by enrichment for a particular cell population by positive selection, or depletion of a particular cell population, by negative selection. In some embodiments, positive or negative selection is accomplished by incubating cells with one or more antibodies or other binding agent that specifically bind to one or more surface markers expressed or expressed (marker+) at a relatively higher level (marker^{high}) on the positively or negatively selected cells, respectively.

[00568] In some embodiments, T cells are separated from a PBMC sample by negative selection of markers expressed on non-T cells, such as B cells, monocytes, or other white blood cells, such as CD14. In some aspects, a CD4+ or CD8+ selection step is used to separate CD4+ helper and CD8+ cytotoxic T-cells. Such CD4+ and CD8+ populations can be further sorted into sub-populations by positive or negative selection for markers expressed or expressed to a relatively higher degree on one or more naive, memory, and/or effector T-cell subpopulations.

[00569] In some embodiments, CD8+ cells are further enriched for or depleted of naive, central memory, effector memory, and/or central memory stem cells, such as by positive or negative selection based on surface antigens associated with the respective subpopulation. In some embodiments, enrichment for central memory T (TCM) cells is carried out to increase efficacy, such as to improve long-term survival, expansion, and/or engraftment following administration, which in some aspects is particularly robust in such sub-populations. See Terakura et al. (2012) Blood. 1:72-82; Wang et al. (2012) J Immunother. 35(9):689-701. In some embodiments, combining TCM-enriched CD8+ T-cells and CD4+ T-cells further enhances efficacy.

[00570] In embodiments, memory T cells are present in both CD62L+ and CD62L- subsets of CD8+ peripheral blood lymphocytes. PBMC can be enriched for or depleted of CD62L-CD8+ and/or CD62L+CD8+ fractions, such as using anti-CD8 and anti-CD62L antibodies.

[00571] In some embodiments, the enrichment for central memory T (TCM) cells is based on positive or high surface expression of CD45RO, CD62L, CCR7, CD28, CD3, and/or CD 127; in some aspects, it is based on negative selection for cells expressing or highly expressing CD45RA

and/or granzyme B. In some aspects, isolation of a CD8+ population enriched for TCM cells is carried out by depletion of cells expressing CD4, CD14, CD45RA, and positive selection or enrichment for cells expressing CD62L. In one aspect, enrichment for central memory T (TCM) cells is carried out starting with a negative fraction of cells selected based on CD4 expression, which is subjected to a negative selection based on expression of CD14 and CD45RA, and a positive selection based on CD62L. Such selections in some aspects are carried out simultaneously and in other aspects are carried out sequentially, in either order. In some aspects, the same CD4 expression-based selection step used in preparing the CD8+ cell population or subpopulation, also is used to generate the CD4+ cell population or sub-population, such that both the positive and negative fractions from the CD4-based separation are retained and used in subsequent steps of the methods, optionally following one or more further positive or negative selection steps.

[00572] In a particular example, a sample of PBMCs or other white blood cell sample is subjected to selection of CD4+ cells, where both the negative and positive fractions are retained. The negative fraction then is subjected to negative selection based on expression of CD14 and CD45RA or ROR1, and positive selection based on a marker characteristic of central memory T-cells, such as CD62L or CCR7, where the positive and negative selections are carried out in either order.

[00573] CD4+ T helper cells are sorted into naive, central memory, and effector cells by identifying cell populations that have cell surface antigens. CD4+ lymphocytes can be obtained by standard methods. In some embodiments, naive CD4+ T lymphocytes are CD45RO-, CD45RA+, CD62L+, CD4+ T-cells. In some embodiments, central memory CD4+ cells are CD62L+ and CD45RO+. In some embodiments, effector CD4+ cells are CD62L- and CD45RO-.

[00574] In one example, to enrich for CD4+ cells by negative selection, a monoclonal antibody cocktail typically includes antibodies to CD14, CD20, CD11b, CD16, HLA-DR, and CD8. In some embodiments, the antibody or binding partner is bound to a solid support or matrix, such as a magnetic bead or paramagnetic bead, to allow for separation of cells for positive and/or negative selection. For example, in some embodiments, the cells and cell populations are separated or isolated using immune-magnetic (or affinity-magnetic) separation techniques (reviewed in Methods in Molecular Medicine, vol. 58: Metastasis Research Protocols, Vol. 2: Cell Behavior In Vitro and In Vivo, p 17-25 Edited by: S. A. Brooks and U. Schumacher Humana Press Inc., Totowa, N.J.).

[00575] In some aspects, the sample or composition of cells to be separated is incubated with small, magnetizable or magnetically responsive material, such as magnetically responsive particles or microparticles, such as paramagnetic beads (e.g., such as Dynabeads or MACS beads). The magnetically responsive material, e.g., particle, generally is directly or indirectly attached to a binding partner, e.g., an antibody, that specifically binds to a molecule, e.g., surface marker, present on the cell, cells, or population of cells that it is desired to separate, e.g., that it is desired to negatively or positively select.

[00576] In some embodiments, the magnetic particle or bead comprises a magnetically responsive material bound to a specific binding member, such as an antibody or other binding partner. There are many well-known magnetically responsive materials used in magnetic separation methods. Suitable magnetic particles include those described in Molday, U.S. Pat. No. 4,452,773, and in European Patent Specification EP 452342 B, which are hereby incorporated by reference. Colloidal sized particles, such as those described in Owen U.S. Pat. No. 4,795,698, and Liberti et al., U.S. Pat. No. 5,200,084 are other examples.

[00577] The incubation generally is carried out under conditions whereby the antibodies or binding partners, or molecules, such as secondary antibodies or other reagents, which specifically bind to such antibodies or binding partners, which are attached to the magnetic particle or bead, specifically bind to cell surface molecules if present on cells within the sample.

[00578] In some aspects, the sample is placed in a magnetic field, and those cells having magnetically responsive or magnetizable particles attached thereto will be attracted to the magnet and separated from the unlabeled cells. For positive selection, cells that are attracted to the magnet are retained; for negative selection, cells that are not attracted (unlabeled cells) are retained. In some aspects, a combination of positive and negative selection is performed during the same selection step, where the positive and negative fractions are retained and further processed or subject to further separation steps.

[00579] In certain embodiments, the magnetically responsive particles are coated in primary antibodies or other binding partners, secondary antibodies, lectins, enzymes, or streptavidin. In certain embodiments, the magnetic particles are attached to cells via a coating of primary antibodies specific for one or more markers. In certain embodiments, the cells, rather than the beads, are labeled with a primary antibody or binding partner, and then cell-type specific secondary antibody- or other binding partner (e.g., streptavidin)-coated magnetic particles, are added. In certain embodiments, streptavidin-coated magnetic particles are used in conjunction with biotinylated primary or secondary antibodies.

[00580] In some embodiments, the magnetically responsive particles are left attached to the cells that are to be subsequently incubated, cultured and/or engineered; in some aspects, the particles are left attached to the cells for administration to a patient. In some embodiments, the magnetizable or magnetically responsive particles are removed from the cells. Methods for removing magnetizable particles from cells are known and include, e.g., the use of competing non-labeled antibodies, magnetizable particles or antibodies conjugated to cleavable linkers, etc. In some embodiments, the magnetizable particles are biodegradable.

[00581] In some embodiments, the affinity-based selection is via magnetic-activated cell sorting (MACS) (Miltenyi Biotech, Auburn, Calif.). Magnetic Activated Cell Sorting (MACS) systems are capable of high-purity selection of cells having magnetized particles attached thereto. In certain embodiments, MACS operates in a mode wherein the non-target and target species are sequentially eluted after the application of the external magnetic field. That is, the cells attached to magnetized particles are held in place while the unattached species are eluted. Then, after this first elution step is completed, the species that were trapped in the magnetic field and were prevented from being eluted are freed in some manner such that they can be eluted and recovered. In certain embodiments, the non-large T cells are labelled and depleted from the heterogeneous population of cells.

[00582] In certain embodiments, the isolation or separation is carried out using a system, device, or apparatus that carries out one or more of the isolation, cell preparation, separation, processing, incubation, culture, and/or formulation steps of the methods. In some aspects, the system is used to carry out each of these steps in a closed or sterile environment, for example, to minimize error, user handling and/or contamination. In one example, the system is a system as described in International Patent Application, Publication Number WO2009/072003, or US 20110003380 A1.

[00583] In some embodiments, the system or apparatus carries out one or more, e.g., all, of the isolation, processing, engineering, and formulation steps in an integrated or self-contained system, and/or in an automated or programmable fashion. In some aspects, the system or apparatus includes a computer and/or computer program in communication with the system or apparatus, which allows a user to program, control, assess the outcome of, and/or adjust various aspects of the processing, isolation, engineering, and formulation steps.

[00584] In some aspects, the separation and/or other steps is carried out using CliniMACS system (Miltenyi Biotic), for example, for automated separation of cells on a clinical-scale level in a closed and sterile system. Components can include an integrated microcomputer, magnetic

separation unit, peristaltic pump, and various pinch valves. The integrated computer in some aspects controls all components of the instrument and directs the system to perform repeated procedures in a standardized sequence. The magnetic separation unit in some aspects includes a movable permanent magnet and a holder for the selection column. The peristaltic pump controls the flow rate throughout the tubing set and, together with the pinch valves, ensures the controlled flow of buffer through the system and continual suspension of cells.

[00585] The CliniMACS system in some aspects uses antibody-coupled magnetizable particles that are supplied in a sterile, non-pyrogenic solution. In some embodiments, after labelling of cells with magnetic particles the cells are washed to remove excess particles. A cell preparation bag is then connected to the tubing set, which in turn is connected to a bag containing buffer and a cell collection bag. The tubing set consists of pre-assembled sterile tubing, including a pre-column and a separation column, and are for single use only. After initiation of the separation program, the system automatically applies the cell sample onto the separation column. Labelled cells are retained within the column, while unlabeled cells are removed by a series of washing steps. In some embodiments, the cell populations for use with the methods described herein are unlabeled and are not retained in the column. In some embodiments, the cell populations for use with the methods described herein are labeled and are retained in the column. In some embodiments, the cell populations for use with the methods described herein are eluted from the column after removal of the magnetic field, and are collected within the cell collection bag.

[00586] In certain embodiments, separation and/or other steps are carried out using the CliniMACS Prodigy system (Miltenyi Biotec). The CliniMACS Prodigy system in some aspects is equipped with a cell processing unity that permits automated washing and fractionation of cells by centrifugation. The CliniMACS Prodigy system can also include an onboard camera and image recognition software that determines the optimal cell fractionation endpoint by discerning the macroscopic layers of the source cell product. For example, peripheral blood may be automatically separated into erythrocytes, white blood cells and plasma layers. The CliniMACS Prodigy system can also include an integrated cell cultivation chamber which accomplishes cell culture protocols such as, e.g., cell differentiation and expansion, antigen loading, and long-term cell culture. Input ports can allow for the sterile removal and replenishment of media and cells can be monitored using an integrated microscope. *See, e.g.,* Klebanoff et al. (2012) J Immunother. 35(9): 651-660, Terakura et al. (2012) Blood. 1:72-82, and Wang et al. (2012) J Immunother. 35(9):689-701.

[00587] In some embodiments, a cell population described herein is collected and enriched (or depleted) via flow cytometry, in which cells stained for multiple cell surface markers are carried in a fluidic stream. In some embodiments, a cell population described herein is collected and enriched (or depleted) via preparative scale (FACS)-sorting. In certain embodiments, a cell population described herein is collected and enriched (or depleted) by use of microelectromechanical systems (MEMS) chips in combination with a FACS-based detection system (see, e.g., WO 2010/033140, Cho et al. (2010) Lab Chip 10, 1567-1573; and Godin et al. (2008) J Biophoton. 1(5):355-376. In both cases, cells can be labeled with multiple markers, allowing for the isolation of well-defined T-cell subsets at high purity.

[00588] In some embodiments, the antibodies or binding partners are labeled with one or more detectable marker, to facilitate separation for positive and/or negative selection. For example, separation may be based on binding to fluorescently labeled antibodies. In some examples, separation of cells based on binding of antibodies or other binding partners specific for one or more cell surface markers are carried in a fluidic stream, such as by fluorescence-activated cell sorting (FACS), including preparative scale (FACS) and/or microelectromechanical systems (MEMS) chips, e.g., in combination with a flow-cytometric detection system. Such methods allow for positive and negative selection based on multiple markers simultaneously.

[00589] In some embodiments, the preparation methods include steps for freezing, e.g., cryopreserving, the cells, either before or after isolation, incubation, and/or engineering. In some embodiments, the freeze and subsequent thaw step removes granulocytes and, to some extent, monocytes in the cell population. In some embodiments, the cells are suspended in a freezing solution, e.g., following a washing step to remove plasma and platelets. Any of a variety of known freezing solutions and parameters in some aspects may be used. One example involves using PBS containing 20% DMSO and 8% human serum albumin (HSA), or other suitable cell freezing media. This can then be diluted 1:1 with media so that the final concentration of DMSO and HSA are 10% and 4%, respectively. Other examples include Cryostor®, CTL-Cryo™ ABC freezing media, and the like. The cells are then frozen to -80 degrees C at a rate of 1degree per minute and stored in the vapor phase of a liquid nitrogen storage tank.

[00590] In some embodiments, the provided methods include cultivation, incubation, culture, and/or genetic engineering steps. For example, in some embodiments, provided are methods for incubating and/or engineering the depleted cell populations and culture-initiating compositions.

[00591] Thus, in some embodiments, the cell populations are incubated in a culture-initiating composition. The incubation and/or engineering may be carried out in a culture vessel, such as a

unit, chamber, well, column, tube, tubing set, valve, vial, culture dish, bag, or other container for culture or cultivating cells.

[00592] In some embodiments, the cells are incubated and/or cultured prior to or in connection with genetic engineering. The incubation steps can include culture, cultivation, stimulation, activation, and/or propagation. In some embodiments, the compositions or cells are incubated in the presence of stimulating conditions or a stimulatory agent. Such conditions include those designed to induce proliferation, expansion, activation, and/or survival of cells in the population, to mimic antigen exposure, and/or to prime the cells for genetic engineering, such as for the introduction of a recombinant antigen receptor.

[00593] The conditions can include one or more of particular media, temperature, oxygen content, carbon dioxide content, time, agents, e.g., nutrients, amino acids, antibiotics, ions, and/or stimulatory factors, such as cytokines, chemokines, antigens, binding partners, fusion proteins, recombinant soluble receptors, and any other agents designed to activate the cells.

[00594] In some embodiments, the stimulating conditions or agents include one or more agent, e.g., ligand, which is capable of activating an intracellular signaling domain of a TCR complex. In some aspects, the agent turns on or initiates TCR/CD3 intracellular signaling cascade in a T-cell. Such agents can include antibodies, such as those specific for a TCR component and/or costimulatory receptor, e.g., anti-CD3, anti-CD28, for example, bound to solid support such as a bead, and/or one or more cytokines. Optionally, the expansion method may further comprise the step of adding anti-CD3 and/or anti CD28 antibody to the culture medium (e.g., at a concentration of at least about 0.5 ng/ml). In some embodiments, the stimulating agents include IL-2 and/or IL-15, for example, an IL-2 concentration of at least about 10 units/mL.

[00595] In some aspects, incubation is carried out in accordance with techniques such as those described in U.S. Pat. No. 6,040,177 to Riddell et al., Klebanoff et al. (2012) J Immunother. 35(9): 651-660, Terakura et al. (2012) Blood. 1:72-82, and/or Wang et al. (2012) J Immunother. 35(9):689-701.

[00596] In some embodiments, the T-cells are expanded by adding to the culture-initiating composition feeder cells, such as non-dividing peripheral blood mononuclear cells (PBMC), (e.g., such that the resulting population of cells contains at least about 5, 10, 20, or 40 or more PBMC feeder cells for each T lymphocyte in the initial population to be expanded); and incubating the culture (e.g. for a time sufficient to expand the numbers of T-cells). In some aspects, the non-dividing feeder cells can comprise gamma-irradiated PBMC feeder cells. In

some embodiments, the PBMC are irradiated with gamma rays in the range of about 3000 to 3600 rads to prevent cell division. In some embodiments, the PBMC feeder cells are inactivated with Mytomicin C. In some aspects, the feeder cells are added to culture medium prior to the addition of the populations of T-cells.

[00597] In some embodiments, the stimulating conditions include temperature suitable for the growth of human T lymphocytes, for example, at least about 25 degrees Celsius, generally at least about 30 degrees, and generally at or about 37 degrees Celsius. Optionally, the incubation may further comprise adding non-dividing EBV-transformed lymphoblastoid cells (LCL) as feeder cells. LCL can be irradiated with gamma rays in the range of about 6000 to 10,000 rads. The LCL feeder cells in some aspects is provided in any suitable amount, such as a ratio of LCL feeder cells to initial T lymphocytes of at least about 10:1.

[00598] In embodiments, antigen-specific T-cells, such as antigen-specific CD4+ and/or CD8+ T-cells, are obtained by stimulating naive or antigen specific T lymphocytes with antigen. For example, antigen-specific T-cell lines or clones can be generated to cytomegalovirus antigens by isolating T-cells from infected subjects and stimulating the cells in vitro with the same antigen.

[00599] In some embodiments, neoantigen-specific T-cells are identified and/or isolated following stimulation with a functional assay (e.g., ELISpot). In some embodiments, neoantigen-specific T-cells are isolated by sorting polyfunctional cells by intracellular cytokine staining. In some embodiments, neoantigen-specific T-cells are identified and/or isolated using activation markers (e.g., CD137, CD38, CD38/HLA-DR double-positive, and/or CD69). In some embodiments, neoantigen-specific CD8+, natural killer T-cells, memory T-cells, and/or CD4+ T-cells are identified and/or isolated using class I or class II multimers and/or activation markers. In some embodiments, neoantigen-specific CD8+ and/or CD4+ T-cells are identified and/or isolated using memory markers (e.g., CD45RA, CD45RO, CCR7, CD27, and/or CD62L). In some embodiments, proliferating cells are identified and/or isolated. In some embodiments, activated T-cells are identified and/or isolated.

[00600] After identification of neoantigen-specific T-cells from a patient sample, the neoantigen-specific TCRs of the identified neoantigen-specific T-cells are sequenced. To sequence a neoantigen-specific TCR, the TCR must first be identified. One method of identifying a neoantigen-specific TCR of a T-cell can include contacting the T-cell with an HLA-multimer (e.g., a tetramer) comprising at least one neoantigen; and identifying the TCR via

binding between the HLA-multimer and the TCR. Another method of identifying a neoantigen-specific TCR can include obtaining one or more T-cells comprising the TCR; activating the one or more T-cells with at least one neoantigen presented on at least one antigen presenting cell (APC); and identifying the TCR via selection of one or more cells activated by interaction with at least one neoantigen.

[00601] After identification of the neoantigen-specific TCR, the TCR can be sequenced. In one embodiment, the methods described above with regard to Section XVII. may be used to sequence TCRs. In another embodiment, TCRa and TCRb of a TCR can be bulk-sequenced and then paired based on frequency. In another embodiment, TCRs can be sequenced and paired using the method of Howie et al., Science Translational Medicine 2015 (doi: 10.1126/scitranslmed.aac5624). In another embodiment, TCRs can be sequenced and paired using the method of Han et al., Nat Biotech 2014 (PMID 24952902, doi 10.1038/nbt.2938). In another embodiment, paired TCR sequences can be obtained using the method described by https://www.biorxiv.org/content/early/2017/05/05/134841 and https://patents.google.com/patent/US20160244825A1/.[158, 159]

[00602] In another embodiment, clonal populations of T cells can be produced by limiting dilution, and then the TCRa and TCRb of the clonal populations of T cells can be sequenced. In yet another embodiment, T-cells can be sorted onto a plate with wells such that there is one T cell per well, and then the TCRa and TCRb of each T cell in each well can be sequenced and paired.

[00603] Next, after neoantigen-specific T-cells are identified from a patient sample and the TCRs of the identified neoantigen-specific T-cells are sequenced, the sequenced TCRs are cloned into new T-cells. These cloned T-cells contain neoantigen-specific receptors, e.g., contain extracellular domains including TCRs. Also provided are populations of such cells, and compositions containing such cells. In some embodiments, compositions or populations are enriched for such cells, such as in which cells expressing the TCRs make up at least 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or more than 99 percent of the total cells in the composition or cells of a certain type such as T-cells or CD8+ or CD4+ cells. In some embodiments, a composition comprises at least one cell containing a TCR disclosed herein. Among the compositions are pharmaceutical compositions and formulations for administration, such as for adoptive cell therapy. Also provided are therapeutic methods for administering the cells and compositions to subjects, e.g., patients.

[00604] Thus also provided are genetically engineered cells expressing TCR(s). The cells generally are eukaryotic cells, such as mammalian cells, and typically are human cells. In some embodiments, the cells are derived from the blood, bone marrow, lymph, or lymphoid organs, are cells of the immune system, such as cells of the innate or adaptive immunity, e.g., myeloid or lymphoid cells, including lymphocytes, typically T-cells and/or NK cells. Other exemplary cells include stem cells, such as multipotent and pluripotent stem cells, including induced pluripotent stem cells (iPSCs). The cells typically are primary cells, such as those isolated directly from a subject and/or isolated from a subject and frozen. In some embodiments, the cells include one or more subsets of T-cells or other cell types, such as whole T-cell populations, CD4+ cells, CD8+ cells, and subpopulations thereof, such as those defined by function, activation state, maturity, potential for differentiation, expansion, recirculation, localization, and/or persistence capacities, antigen-specificity, type of antigen receptor, presence in a particular organ or compartment, marker or cytokine secretion profile, and/or degree of differentiation. With reference to the subject to be treated, the cells may be allogeneic and/or autologous. Among the methods include off-the-shelf methods. In some aspects, such as for off-the-shelf technologies, the cells are pluripotent and/or multipotent, such as stem cells, such as induced pluripotent stem cells (iPSCs). In some embodiments, the methods include isolating cells from the subject, preparing, processing, culturing, and/or engineering them, as described herein, and re-introducing them into the same patient, before or after cryopreservation.

[00605] Among the sub-types and subpopulations of T-cells and/or of CD4+ and/or of CD8+ T-cells are naive T (TN) cells, effector T-cells (TEFF), memory T-cells and sub-types thereof, such as stem cell memory T (TSCM), central memory T (TCM), effector memory T (TEM), or terminally differentiated effector memory T-cells, tumor-infiltrating lymphocytes (TIL), immature T-cells, mature T-cells, helper T-cells, cytotoxic T-cells, mucosa-associated invariant T (MALT) cells, naturally occurring and adaptive regulatory T (Treg) cells, helper T-cells, such as TH1 cells, TH2 cells, TH3 cells, TH17 cells, TH9 cells, TH22 cells, follicular helper T-cells, alpha/beta T-cells, and delta/gamma T-cells.

[00606] In some embodiments, the cells are natural killer (NK) cells. In some embodiments, the cells are monocytes or granulocytes, e.g., myeloid cells, macrophages, neutrophils, dendritic cells, mast cells, eosinophils, and/or basophils.

[00607] The cells may be genetically modified to reduce expression or knock out endogenous TCRs. Such modifications are described in Mol Ther Nucleic Acids. 2012 Dec;

1(12): e63; Blood. 2011 Aug 11;118(6):1495-503; Blood. 2012 Jun 14; 119(24): 5697–5705; Torikai, Hiroki et al "HLA and TCR Knockout by Zinc Finger Nucleases: Toward "off-the-Shelf' Allogeneic T-Cell Therapy for CD19+ Malignancies.." Blood 116.21 (2010): 3766; Blood. 2018 Jan 18;131(3):311-322. doi: 10.1182/blood-2017-05-787598; and WO2016069283, which are incorporated by reference in their entirety.

[00608] The cells may be genetically modified to promote cytokine secretion. Such modifications are described in Hsu C, Hughes MS, Zheng Z, Bray RB, Rosenberg SA, Morgan RA. Primary human T lymphocytes engineered with a codon-optimized IL-15 gene resist cytokine withdrawal-induced apoptosis and persist long-term in the absence of exogenous cytokine. J Immunol. 2005;175:7226–34; Quintarelli C, Vera JF, Savoldo B, Giordano Attianese GM, Pule M, Foster AE, Co-expression of cytokine and suicide genes to enhance the activity and safety of tumor-specific cytotoxic T lymphocytes. Blood. 2007;110:2793–802; and Hsu C, Jones SA, Cohen CJ, Zheng Z, Kerstann K, Zhou J, Cytokine-independent growth and clonal expansion of a primary human CD8+ T-cell clone following retroviral transduction with the IL-15 gene. Blood. 2007;109:5168–77.

[00609] Mismatching of chemokine receptors on T-cells and tumor-secreted chemokines has been shown to account for the suboptimal trafficking of T-cells into the tumor microenvironment. To improve efficacy of therapy, the cells may be genetically modified to increase recognition of chemokines in tumor micro environment. Examples of such modifications are described in Moon, EKCarpenito, CSun, JWang, LCKapoor, VPredina, J Expression of a functional CCR2 receptor enhances tumor localization and tumor eradication by retargeted human T-cells expressing a mesothelin-specific chimeric antibody receptor.Clin Cancer Res. 2011; 17: 4719-4730; and.Craddock, JALu, ABear, APule, MBrenner, MKRooney, CM et al. Enhanced tumor trafficking of GD2 chimeric antigen receptor T-cells by expression of the chemokine receptor CCR2b.J Immunother. 2010; 33: 780-788.

[00610] The cells may be genetically modified to enhance expression of costimulatory/enhancing receptors, such as CD28 and 41BB.

[00611] Adverse effects of T-cell therapy can include cytokine release syndrome and prolonged B-cell depletion. Introduction of a suicide/safety switch in the recipient cells may improve the safety profile of a cell-based therapy. Accordingly, the cells may be genetically modified to include a suicide/safety switch. The suicide/safety switch may be a gene that confers sensitivity to an agent, e.g., a drug, upon the cell in which the gene is expressed, and which causes the cell to die when the cell is contacted with or exposed to the agent.

Exemplary suicide/safety switches are described in Protein Cell. 2017 Aug; 8(8): 573–589. The suicide/safety switch may be HSV-TK. The suicide/safety switch may be cytosine daminase, purine nucleoside phosphorylase, or nitroreductase. The suicide/safety switch may be RapaCIDe™, described in U.S. Patent Application Pub. No. US20170166877A1. The suicide/safety switch system may be CD20/Rituximab, described in Haematologica. 2009 Sep; 94(9): 1316–1320. These references are incorporated by reference in their entirety.

[00612] The TCR may be introduced into the recipient cell as a split receptor which assembles only in the presence of a heterodimerizing small molecule. Such systems are described in Science. 2015 Oct 16; 350(6258): aab4077, and in U.S. Patent No. 9,587,020, which are hereby incorporated by reference.

[00613] In some embodiments, the cells include one or more nucleic acids, e.g., a polynucleotide encoding a TCR disclosed herein, wherein the polynucleotide is introduced via genetic engineering, and thereby express recombinant or genetically engineered TCRs as disclosed herein. In some embodiments, the nucleic acids are heterologous, i.e., normally not present in a cell or sample obtained from the cell, such as one obtained from another organism or cell, which for example, is not ordinarily found in the cell being engineered and/or an organism from which such cell is derived. In some embodiments, the nucleic acids are not naturally occurring, such as a nucleic acid not found in nature, including one comprising chimeric combinations of nucleic acids encoding various domains from multiple different cell types.

[00614] The nucleic acids may include a codon-optimized nucleotide sequence. Without being bound to a particular theory or mechanism, it is believed that codon optimization of the nucleotide sequence increases the translation efficiency of the mRNA transcripts. Codon optimization of the nucleotide sequence may involve substituting a native codon for another codon that encodes the same amino acid, but can be translated by tRNA that is more readily available within a cell, thus increasing translation efficiency. Optimization of the nucleotide sequence may also reduce secondary mRNA structures that would interfere with translation, thus increasing translation efficiency.

[00615] A construct or vector may be used to introduce the TCR into the recipient cell. Exemplary constructs are described herein. Polynucleotides encoding the alpha and beta chains of the TCR may in a single construct or in separate constructs. The polynucleotides encoding the alpha and beta chains may be operably linked to a promoter, e.g., a heterologous promoter. The heterologous promoter may be a strong promoter, e.g.,

EF1alpha, CMV, PGK1, Ubc, beta actin, CAG promoter, and the like. The heterologous promoter may be a weak promoter. The heterologous promoter may be an inducible promoter. Exemplary inducible promoters include, but are not limited to TRE, NFAT, GAL4, LAC, and the like. Other exemplary inducible expression systems are described in U.S. Patent Nos. 5,514,578; 6,245,531; 7,091,038 and European Patent No. 0517805, which are incorporated by reference in their entirety.

[00616] The construct for introducing the TCR into the recipient cell may also comprise a polynucleotide encoding a signal peptide (signal peptide element). The signal peptide may promote surface trafficking of the introduced TCR. Exemplary signal peptides include, but are not limited to CD8 signal peptide, immunoglobulin signal peptides, where specific examples include GM-CSF and IgG kappa. Such signal peptides are described in Trends Biochem Sci. 2006 Oct;31(10):563-71. Epub 2006 Aug 21; and An, et al. "Construction of a New Anti-CD19 Chimeric Antigen Receptor and the Anti-Leukemia Function Study of the Transduced T-cells." Oncotarget 7.9 (2016): 10638–10649. PMC. Web. 16 Aug. 2018; which are hereby incorporated by reference.

[00617] In some cases, e.g., cases where the alpha and beta chains are expressed from a single construct or open reading frame, or cases wherein a marker gene is included in the construct, the construct may comprise a ribosomal skip sequence. The ribosomal skip sequence may be a 2A peptide, e.g., a P2A or T2A peptide. Exemplary P2A and T2A peptides are described in Scientific Reports volume 7, Article number: 2193 (2017), hereby incorporated by reference in its entirety. In some cases, a FURIN/PACE cleavage site is introduced upstream of the 2A element. FURIN/PACE cleavage sites are described in, e.g., http://www.nuolan.net/substrates.html. The cleavage peptide may also be a factor Xa cleavage site. In cases where the alpha and beta chains are expressed from a single construct or open reading frame, the construct may comprise an internal ribosome entry site (IRES).

[00618] The construct may further comprise one or more marker genes. Exemplary marker genes include but are not limited to GFP, luciferase, HA, lacZ. The marker may be a selectable marker, such as an antibiotic resistance marker, a heavy metal resistance marker, or a biocide resistant marker, as is known to those of skill in the art. The marker may be a complementation marker for use in an auxotrophic host. Exemplary complementation markers and auxotrophic hosts are described in Gene. 2001 Jan 24;263(1-2):159-69. Such markers may be expressed via an IRES, a frameshift sequence, a 2A peptide linker, a fusion with the TCR, or expressed separately from a separate promoter.

[00619] Exemplary vectors or systems for introducing TCRs into recipient cells include, but are not limited to Adeno-associated virus, Adenovirus, Adenovirus + Modified vaccinia, Ankara virus (MVA), Adenovirus + Retrovirus, Adenovirus + Sendai virus, Adenovirus + Vaccinia virus, Alphavirus (VEE) Replicon Vaccine, Antisense oligonucleotide, Bifidobacterium longum, CRISPR-Cas9, E. coli, Flavivirus, Gene gun, Herpesviruses, Herpes simplex virus, Lactococcus lactis, Electroporation, Lentivirus, Lipofection, Listeria monocytogenes, Measles virus, Modified Vaccinia Ankara virus (MVA), mRNA Electroporation, Naked/Plasmid DNA, Naked/Plasmid DNA + Adenovirus, Naked/Plasmid DNA + Modified Vaccinia Ankara virus (MVA), Naked/Plasmid DNA + RNA transfer, Naked/Plasmid DNA + Vaccinia virus, Naked/Plasmid DNA + Vesicular stomatitis virus, Newcastle disease virus, Non-viral, PiggyBac™ (PB) Transposon, nanoparticle-based systems, Poliovirus, Poxvirus, Poxvirus + Vaccinia virus, Retrovirus, RNA transfer, RNA transfer + Naked/Plasmid DNA, RNA virus, Saccharomyces cerevisiae, Salmonella typhimurium, Semliki forest virus, Sendai virus, Shigella dysenteriae, Simian virus, siRNA, Sleeping Beauty transposon, Streptococcus mutans, Vaccinia virus, Venezuelan equine encephalitis virus replicon, Vesicular stomatitis virus, and Vibrio cholera.

[00620] In preferred embodiments, the TCR is introduced into the recipient cell via adeno associated virus (AAV), adenovirus, CRISPR-CAS9, herpesvirus, lentivirus, lipofection, mRNA electroporation, PiggyBac™ (PB) Transposon, retrovirus, RNA transfer, or Sleeping Beauty transposon.

[00621] In some embodiments, a vector for introducing a TCR into a recipient cell is a viral vector. Exemplary viral vectors include adenoviral vectors, adeno-associated viral (AAV) vectors, lentiviral vectors, herpes viral vectors, retroviral vectors, and the like. Such vectors are described herein.

[00622] Exemplary embodiments of TCR constructs for introducing a TCR into recipient cells is shown in FIG. 33. In some embodiments, a TCR construct includes, from the 5'-3' direction, the following polynucleotide sequences: a promoter sequence, a signal peptide sequence, a TCR β variable (TCRβv) sequence, a TCR β constant (TCRβc) sequence, a cleavage peptide (e.g., P2A), a signal peptide sequence, a TCR α variable (TCRαv) sequence, and a TCR α constant (TCRαc) sequence. In some embodiments, the TCRβc and TCRαc sequences of the construct include one or more murine regions, e.g., full murine constant sequences or human ➔ murine amino acid exchanges as described herein. In some embodiments, the construct further includes, 3' of the TCRαc sequence, a cleavage peptide

sequence (e.g., T2A) followed by a reporter gene. In an embodiment, the construct includes, from the 5'-3' direction, the following polynucleotide sequences: a promoter sequence, a signal peptide sequence, a TCR β variable (TCRβv) sequence, a TCR β constant ((TCRβc) sequence containing one or more murine regions, a cleavage peptide (e.g., P2A), a signal peptide sequence, a TCR α variable (TCRαv) sequence, and a TCR α constant (TCRαc) sequence containing one or more murine regions, a cleavage peptide (e.g., T2A), and a reporter gene.

[00623] FIG. 34 depicts an exemplary P526 construct backbone nucleotide sequence for cloning TCRs into expression systems for therapy development.

[00624] FIG. 35 depicts an exemplary construct sequence for cloning patient neoantigen-specific TCR, clonotype 1 into expression systems for therapy development.

[00625] FIG. 36 depicts an exemplary construct sequence for cloning patient neoantigen-specific TCR, clonotype 3 into expression systems for therapy development.

[00626] Also provided are isolated nucleic acids encoding TCRs, vectors comprising the nucleic acids, and host cells comprising the vectors and nucleic acids, as well as recombinant techniques for the production of the TCRs.

[00627] The nucleic acids may be recombinant. The recombinant nucleic acids may be constructed outside living cells by joining natural or synthetic nucleic acid segments to nucleic acid molecules that can replicate in a living cell, or replication products thereof. For purposes herein, the replication can be in vitro replication or in vivo replication.

[00628] For recombinant production of a TCR, the nucleic acid(s) encoding it may be isolated and inserted into a replicable vector for further cloning (i.e., amplification of the DNA) or expression. In some aspects, the nucleic acid may be produced by homologous recombination, for example as described in U.S. Patent No. 5,204,244, incorporated by reference in its entirety.

[00629] Many different vectors are known in the art. The vector components generally include one or more of the following: a signal sequence, an origin of replication, one or more marker genes, an enhancer element, a promoter, and a transcription termination sequence, for example as described in U.S. Patent No. 5,534,615, incorporated by reference in its entirety.

[00630] Exemplary vectors or constructs suitable for expressing a TCR, antibody, or antigen binding fragment thereof, include, e.g., the pUC series (Fermentas Life Sciences), the pBluescript series (Stratagene, LaJolla, CA), the pET series (Novagen, Madison, WI), the pGEX series (Pharmacia Biotech, Uppsala, Sweden), and the pEX series (Clontech, Palo

Alto, CA). Bacteriophage vectors, such as AGTlO, AGTl 1, AZapII (Stratagene), AEMBL4, and ANMl 149, are also suitable for expressing a TCR disclosed herein.

XIX. Treatment Overview Flow Chart

[00631] FIG. 37 is a flow chart of a method for providing a customized, neoantigen-specific treatment to a patient, in accordance with an embodiment. In other embodiments, the method may include different and/or additional steps than those shown in FIG. 37. Additionally, steps of the method may be performed in different orders than the order described in conjunction with FIG. 37 in various embodiments.

[00632] The presentation models are trained 3701 using mass spectrometry data as described above. A patient sample is obtained 3702. In some embodiments, the patient sample comprises a tumor biopsy and/or the patient's peripheral blood. The patient sample obtained in step 3702 is sequenced to identify data to input into the presentation models to predict the likelihoods that tumor antigen peptides from the patient sample will be presented. Presentation likelihoods of tumor antigen peptides from the patient sample obtained in step 3702 are predicted 3703 using the trained presentation models. Treatment neoantigens are identified 3704 for the patient based on the predicted presentation likelihoods. Next, another patient sample is obtained 3705. The patient sample may comprise the patient's peripheral blood, tumor-infiltrating lymphocytes (TIL), lymph, lymph node cells, and/or any other source of T-cells. The patient sample obtained in step 3705 is screened 3706 *in vivo* for neoantigen-specific T-cells.

[00633] At this point in the treatment process, the patient can either receive T-cell therapy and/or a vaccine treatment. To receive a vaccine treatment, the neoantigens to which the patient's T-cells are specific are identified 3714. Then, a vaccine including the identified neoantigens is created 3715. Finally, the vaccine is administered 3716 to the patient.

[00634] To receive T-cell therapy, the neoantigen-specific T-cells undergo expansion and/or new neoantigen-specific T-cells are genetically engineered. To expand the neoantigen-specific T-cells for use in T-cell therapy, the cells are simply expanded 3707 and infused 3708 into the patient.

[00635] To genetically engineer new neoantigen-specific T-cells for T-cell therapy, the TCRs of the neoantigen-specific T-cells that were identified *in vivo* are sequenced 3709. Next, these TCR sequences are cloned 3710 into an expression vector. The expression vector

3710 is then transfected 3711 into new T-cells. The transfected T-cells are 3712 expanded. And finally, the expanded T-cells are infused 3713 into the patient.

[00636] A patient may receive both T-cell therapy and vaccine therapy. In one embodiment, the patient first receives vaccine therapy then receives T-cell therapy. One advantage of this approach is that the vaccine therapy may increase the number of tumor-specific T-cells and the number of neoantigens recognized by detectable levels of T-cells.

[00637] In another embodiment, a patient may receive T-cell therapy followed by vaccine therapy, wherein the set of epitopes included in the vaccine comprises one or more of the epitopes targeted by the T-cell therapy. One advantage of this approach is that administration of the vaccine may promote expansion and persistence of the therapeutic T-cells.

XX. Example Computer

[00638] FIG. 38 illustrates an example computer 3800 for implementing the entities shown in FIGS. 1 and 3. The computer 3800 includes at least one processor 3802 coupled to a chipset 3804. The chipset 3804 includes a memory controller hub 3820 and an input/output (I/O) controller hub 3822. A memory 3806 and a graphics adapter 3812 are coupled to the memory controller hub 3820, and a display 3818 is coupled to the graphics adapter 3812. A storage device 3808, an input device 3814, and network adapter 3816 are coupled to the I/O controller hub 3822. Other embodiments of the computer 3800 have different architectures.

[00639] The storage device 3808 is a non-transitory computer-readable storage medium such as a hard drive, compact disk read-only memory (CD-ROM), DVD, or a solid-state memory device. The memory 3806 holds instructions and data used by the processor 3802. The input interface 3814 is a touch-screen interface, a mouse, track ball, or other type of pointing device, a keyboard, or some combination thereof, and is used to input data into the computer 3800. In some embodiments, the computer 3800 may be configured to receive input (e.g., commands) from the input interface 3814 via gestures from the user. The graphics adapter 3812 displays images and other information on the display 3818. The network adapter 3816 couples the computer 3800 to one or more computer networks.

[00640] The computer 3800 is adapted to execute computer program modules for providing functionality described herein. As used herein, the term "module" refers to computer program logic used to provide the specified functionality. Thus, a module can be implemented in hardware, firmware, and/or software. In one embodiment, program modules are stored on the storage device 3808, loaded into the memory 3806, and executed by the processor 3802.

[00641] The types of computers 3800 used by the entities of FIG. 1 can vary depending upon the embodiment and the processing power required by the entity. For example, the presentation identification system 160 can run in a single computer 3800 or multiple computers 3800 communicating with each other through a network such as in a server farm. The computers 3800 can lack some of the components described above, such as graphics adapters 3812, and displays 3818.

Supplementary Table 1
Demographics of NSCLS Patients

Patient ID	Age Range (Years)	Gender	Race	Year of Initial (Lung Cancer) Diagnosis	Tumor Stage (At Enrollment)	Location of Primary Tumor	Histological Type
1-001-002	81-90	Male	White	2010	IIIB	Lung	Non-squamous
1-024-001	81-90	Male	White	2016	IV	Lung	Sarcomatoid pulmonary carcinoma
1-024-002	51-60	Female	White	2016	IV	Lung	Adenocarcinoma
1-038-001	61-70	Male	White	2016	IV	Lung	Adenocarcinoma

Supplementary Table 1
Demographics of NSCLS Patients

Systemic NSCLC-Directed Therapy	Current Anti-PD(L)-1 Therapy	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C	Expressed Mutations
Carboplantin	Nivolumab	A*01:01	A*01:01	B*08:01	B*51:01	C*01:02	C*07:01	122
	Pembrolizumab	A*32:01	A*03:01	B*27:05	B*27:05	C*02:02	C*02:02	83
DOCEtaxel, Bevacizumab, Ramucirumab, Pemetrexed Disodium	Nivolumab	A*68:01	A*68:01	B*40:02	B*40:27	C*03:04	C*03:04	38
premetexed, Cisplatin	Nivolumab	A*69:01	A*01:02	B*41:01	B*49:01	C*17:01	C*07:01	158

Supplementary Table 1
Demographics of NSCLS Patients

Nonsynonymous Mutations	Normal DNA Median Exon Coverage	Tumor DNA Median Exon Coverage	RNA PF Unique Reads (M)	Known Drivers	Likely Drivers	Median VAF
232	145	552	173	KRAS_G12D, TP53_R213*	STK11_G52fs	0.22
143	165	508	131.9	KRAS_G12C, TP53_R280T	PML_E43*, NF2_R341*	0.093
69	190	454	114.4	KRAS_G12S, TP53_Q331*	STK11_E199*	0.182
265	158	983	311.8	KRAS_G12V	KDM5C_E303*	0.19

Supplementary Table 1
Demographics of NSCLS Patients

Patient ID	Age Range (Years)	Gender	Race	Year of Initial (Lung Cancer) Diagnosis	Tumor Stage (At Enrollment)	Location of Primary Tumor	Histological Type
1-050-001	71-80	Female	White	2015	IIIB	Lung	Adenocarcinoma
CU05	71-80	Female	White	2013	IV	Lung	Lung Squamous
CU04	61-70	Female	Hispanic or Latino	2013	I	Lung	Adenocarcinoma
CU03	61-70	Male	African American	2016	I	Lung	Lung Squamous
CU02	61-70	Male	White	2016	I	Lung	Lung Squamous

Supplementary Table 1
Demographics of NSCLS Patients

Systemic NSCLC-Directed Therapy	Current Anti-PD(L)-1 Therapy	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C	Expressed Mutations
ETOPOSIDE, cisplatin	Nivolumab	A*29:02	A*26:01	B*44:03	B*07:05	C*16:01	C*15:05	53
carboplatin plus pemetrexed	Nivolumab	A*24:02	A*68:02	B*14:02	B*15:17	C*07:01	C*08:02	65
	durvalumab plus tremelimumab	A*24:26	A*26:01	B*18:01	B*38:01	C*12:03	C*12:03	336
	n/a	A*23:01	A*01:01	B*08:01	B*15:03	C*01:02	C*12:03	105
carboplatin + gemcitabine	n/a	A*02:01	A*03:01	B*07:02	B*57:01	C*07:02	C*06:02	102

Supplementary Table 1
Demographics of NSCLS Patients

Nonsynonymous Mutations	Normal DNA Median Exon Coverage	Tumor DNA Median Exon Coverage	RNA PF Unique Reads (M)	Known Drivers	Likely Drivers	Median VAF
92	117	556	119			0.059
109	191	448	83.6			0.095
511	213	552	240.4	TP53_R158G	NFKBIE_G41fs, CDH1_Q346*, NF1_D2163fs, MED12_R730*	0.224
187	114	830	182.1			0.242
174	105	738	185.3	TP53_R175H	ATR_Q195*	0.32

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Patient	Peptide	Individual Peptide Response (Any Time Point)	Individual Pepetide Response Notes	Pool ID	Pool Response (Any Time Point)	Mutation
1-001-002	HSPFTATSL	N		1-001-002_pool_1	N	chr15_28215653_C_A
1-001-002	DPEEVLVTV	N		1-001-002_pool_1	N	chr17_59680958_C_T
1-001-002	ELDPDIQLEY	N		1-001-002_pool_1	N	chr13_30210371_C_A
1-001-002	TPLTKDVTL	N		1-001-002_pool_1	N	chr5_78100974_A_T
1-001-002	DGVGKSAL	N		1-001-002_pool_1	N	chr12_25245350_C_T
1-001-002	YTTVRALTL	N		1-001-002_pool_1	N	chr17_28339664_G_T
1-001-002	TPSAAVKLI	N		1-001-002_pool_1	N	chr15_81319417_T_C
1-001-002	WPVLLLNV	N		1-001-002_pool_1	N	chr3_179025167_AAC_A
1-001-002	ELNARRCSF	N		1-001-002_pool_1	N	chr18_79943341_G_A

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Mutation Type	Gene	Protein Effect	TPM	Most Probable Restriction covered by Full MS Model	Full MS Model Rank	MHCFlurry Rank	MHCFlurry (nM)	Most Probable Restriction covered by MHCFlurry
snp	HERC2	A2060S	41.9	HLA-C*01:02	0	95	5169.68205	FALSE
snp	CLTC	S989L	272.1	HLA-B*51:01	1	61	3455.25069	TRUE
snp	KATNAL1	D407Y	12.81	HLA-A*01:01	2	1	24.2177849	TRUE
snp	AP3B1	S817T	44.4	HLA-B*08:01	3	2	48.9740194	TRUE
snp	KRAS	G12D	40.75	HLA-B*08:01	4	89	4714.29522	TRUE
snp	TNFAIP1	R48L	45.62	HLA-B*08:01	5	26	973.417701	TRUE
snp	STARD5	M108V	1.95	HLA-B*51:01	6	39	2030.48603	TRUE
del_fs	ZMAT3	V240fs	14.99	HLA-B*51:01	7	16	600.564752	TRUE
snp	PQLC1	R109C	33.89	HLA-B*08:01	8	5	62.0439997	TRUE

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Patient	Peptide	Individual Peptide Response (Any Time Point)	Individual Pepetide Response Notes	Pool ID	Pool Response (Any Time Point)	Mutation
1-001-002	QMKNPILEL	N		1-001-002_pool_1	N	chr9_127663287_G_T
1-001-002	LTEKVSLLK	N		1-001-002_pool_2	N	chr9_92719180_C_T
1-001-002	SPFTATSL	N		1-001-002_pool_2	N	chr15_28215653_C_A
1-001-002	NVDMRTISF	N		1-001-002_pool_2	N	chr9_121353262_T_A
1-001-002	TSIVVSQTL	N		1-001-002_pool_2	N	chr4_39205691_C_T
1-001-002	HIKIEPVAI	N		1-001-002_pool_2	N	chr13_73062087_C_T
1-001-002	DSPDGSNGL	N		1-001-002_pool_2	N	chr20_44197575_C_T
1-001-002	YTAVHYAASY	N		1-001-002_pool_2	N	chr12_56248788_C_A
1-001-002	VGADGVGKSA L	N		1-001-002_pool_2	N	chr12_25245350_C_T

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Mutation Type	Gene	Protein Effect	TPM	Most Probable Restriction covered by Full MS Model	Full MS Model Rank	MHCFlurry Rank	MHCFlurry (nM)	Most Probable Restriction covered by MHCFlurry
snp	STXBP1	R171L	38.76	HLA-B*08:01	9	20	674.64733	TRUE
snp	BICD2	E489K	42.66	HLA-A*01:01	10	10	428.744925	TRUE
snp	HERC2	A2060S	41.9	HLA-B*08:01	11	4	59.1155419	TRUE
snp	STOM	K93N	360.6	HLA-B*08:01	12	30	1490.72261	TRUE
snp	WDR19	A282V	18.12	HLA-B*08:01	13	176	9862.33009	TRUE
snp	KLF5	T163I	25.77	HLA-B*08:01	14	27	1122.27455	TRUE
snp	OSER1	S119N	20.7	HLA-C*01:02	15	471	21598.414	FALSE
snp	ANKRD52	A559S	18.32	HLA-A*01:01	16	0	11.5906737	TRUE
snp	KRAS	G12D	40.75	HLA-C*01:02	17	370	17985.3612	FALSE

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Patient	Peptide	Individual Peptide Response (Any Time Point)	Individual Pepetide Response Notes	Pool ID	Pool Response (Any Time Point)	Mutation
1-001-002	MMPPLPGI	N		1-001-002_pool_2	N	chr17_32369404_A_T
1-001-002	FPYPGMTNQ	N		1-001-002_pool_2	N	chr5_109186272_G_T
1-024-001	VTNHAPLSW	N		1-024-001_pool_1	Y	chr3_125552370_C_A
1-024-001	GTKKDVDVLK	Y		1-024-001_pool_1	Y	chr20_56513366_G_A
1-024-001	GLNVPVQSNK	N		1-024-001_pool_1	Y	chr4_88390868_G_T
1-024-001	VVVGACGVGK	N		1-024-001_pool_1	Y	chr12_25245351_C_A
1-024-001	AQFAGKDQTY	N		1-024-001_pool_1	Y	chr9_89045819_C_A
1-024-001	KVVLPSDVTSY	N		1-024-001_pool_1	Y	chr3_48591778_G_T
1-024-001	MLMKNISTK	N		1-024-001_pool_1	Y	chr12_6959976_G_A

Supplementary Table 2

Peptides Tested for T-Cell Recognition in NSCLC Patients

Mutation Type	Gene	Protein Effect	TPM	Most Probable Restriction covered by Full MS Model	Full MS Model Rank	MHCFlurry Rank	MHCFlurry (nM)	Most Probable Restriction covered by MHCFlurry
snp	ZNF207	Q409L	186	HLA-B*51:01	18	136	7609.76602	TRUE
snp	FER	C759F	67.36	HLA-B*51:01	19	38	1999.07208	TRUE
snp	OSBPL1A	G489W	24.12	HLA-A*32:01	0	7	77.009026	TRUE
snp	RTFDC1	E177K	61.32	HLA-A*03:01	1	70	2168.51668	TRUE
snp	HERC6	R218L	8.7	HLA-A*03:01	2	4	59.675168	TRUE
snp	KRAS	G12C	40.05	HLA-A*03:01	3	11	133.648023	TRUE
snp	SHC3	E376D	8.88	HLA-A*32:01	4	91	3715.42819	TRUE
snp	COL7A1	R468S	25.42	HLA-A*32:01	6	85	3234.15772	TRUE
snp	PTPN6	E471K	105.4	HLA-A*03:01	7	0	12.2301919	TRUE

DEMANDE OU BREVET VOLUMINEUX

LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVET COMPREND PLUS D'UN TOME.

CECI EST LE TOME 1 DE 2
CONTENANT LES PAGES 1 À 165

NOTE : Pour les tomes additionels, veuillez contacter le Bureau canadien des brevets

JUMBO APPLICATIONS/PATENTS

THIS SECTION OF THE APPLICATION/PATENT CONTAINS MORE THAN ONE VOLUME

THIS IS VOLUME 1 OF 2
CONTAINING PAGES 1 TO 165

NOTE: For additional volumes, please contact the Canadian Patent Office

NOM DU FICHIER / FILE NAME :

NOTE POUR LE TOME / VOLUME NOTE:

CLAIMS

1. A method for identifying at least one neoantigen from one or more tumor cells of a subject that are likely to be presented by one or more MHC alleles on a surface of the tumor cells, the method comprising the steps of:

obtaining at least one of exome, transcriptome, or whole genome nucleotide sequencing data from the tumor cells and normal cells of the subject, wherein the nucleotide sequencing data is used to obtain data representing peptide sequences of each of a set of neoantigens identified by comparing the nucleotide sequencing data from the tumor cells and the nucleotide sequencing data from the normal cells, wherein the peptide sequence of each neoantigen comprises at least one alteration that makes it distinct from the corresponding wild-type peptide sequence identified from the normal cells of the subject;

encoding the peptide sequences of each of the neoantigens into a corresponding numerical vector, each numerical vector including information regarding a plurality of amino acids that make up the peptide sequence and a set of positions of the amino acids in the peptide sequence;

obtaining at least one of exome, transcriptome or whole genome nucleotide sequencing data from the tumor cells the subject, wherein the nucleotide sequencing data is used to obtain data representing a peptide sequence of each of the one or more MHC alleles of the subject;

encoding the peptide sequences of each of the one or more MHC alleles of the subject into a corresponding numerical vector, each numerical vector including information regarding a plurality of amino acids that make up the peptide sequence and a set of positions of the amino acids in the peptide sequence;

inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles, using a computer processor, into a machine-learned presentation model to generate a set of presentation likelihoods for the set of neoantigens, each presentation likelihood in the set representing the likelihood that a corresponding neoantigen is presented by the one or more MHC alleles on the surface of the tumor cells of the subject, the machine-learned presentation model comprising:

a plurality of parameters identified at least based on a training data set comprising:

for each sample in a plurality of samples, a label obtained by mass spectrometry measuring presence of peptides bound to at least one MHC allele in a set of MHC alleles identified as present in the sample;

for each of the samples, training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the peptides and a set of positions of the amino acids in the peptides; and

for each of the samples, training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the at least one MHC allele bound to the peptides of the sample and a set of positions of the amino acids in the at least one MHC allele;

a function representing a relation between the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles received as input, and the presentation likelihood generated as output based on the numerical vectors and the parameters;

selecting a subset of the set of neoantigens based on the set of presentation likelihoods to generate a set of selected neoantigens; and

returning the set of selected neoantigens.

2. The method of claim 1, wherein inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model comprises:

applying the machine-learned presentation model to the peptide sequence of the neoantigen and to the peptide sequence of the one or more MHC alleles to generate a dependency score for each of the one or more MHC alleles indicating whether the MHC allele will present the neoantigen based on the particular amino acids at the particular positions of the peptide sequences.

3. The method of claim 2, wherein inputting the numerical vectors encoding the peptide

sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model further comprises:

transforming the dependency scores to generate a corresponding per-allele likelihood for each MHC allele indicating a likelihood that the corresponding MHC allele will present the corresponding neoantigen; and

combining the per-allele likelihoods to generate the presentation likelihood of the neoantigen.

4. The method of claim 3, wherein transforming the dependency scores models the presentation of the neoantigen as mutually exclusive across the one or more MHC alleles.

5. The method of claim 2, wherein inputting the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles into the machine-learned presentation model further comprises:

transforming a combination of the dependency scores to generate the presentation likelihood, wherein transforming the combination of the dependency scores models the presentation of the neoantigen as interfering between the one or more MHC alleles.

6. The method of any one of claims 2-5, wherein the set of presentation likelihoods are further identified by at least one or more allele noninteracting features, and further comprising:

applying the machine-learned presentation model to the allele noninteracting features to generate a dependency score for the allele noninteracting features indicating whether the peptide sequence of the corresponding neoantigen will be presented based on the allele noninteracting features.

7. The method of claim 6, further comprising:

combining the dependency score for each MHC allele of the one or more MHC alleles with the dependency score for the allele noninteracting features;

transforming the combined dependency scores for each MHC allele to generate a per-allele likelihood for each MHC allele indicating a likelihood that the corresponding MHC allele will present the corresponding neoantigen; and

combining the per-allele likelihoods to generate the presentation likelihood.

8. The method of claim 6, further comprising:

combining the dependency scores for each of the MHC alleles and the dependency score for the allele noninteracting features; and

transforming the combined dependency scores to generate the presentation likelihood.

9. The method of any one of claims 1-8, wherein the one or more MHC alleles include two or more different MHC alleles.

10. The method of any one of claims 1-9, wherein the peptide sequences comprise peptide sequences having lengths other than 9 amino acids.

11. The method of any one of claims 1-10, wherein encoding a peptide sequence comprises encoding the peptide sequence using a one-hot encoding scheme.

12. The method of any one of claims 1-11, wherein the plurality of samples comprise at least one of:

(a) one or more cell lines engineered to express a single MHC allele;

(b) one or more cell lines engineered to express a plurality of MHC alleles;

(c) one or more human cell lines obtained or derived from a plurality of patients;

(d) fresh or frozen tumor samples obtained from a plurality of patients; and

(e) fresh or frozen tissue samples obtained from a plurality of patients.

13. The method of any one of claims 1-12, wherein the training data set further comprises at least one of:

(a) data associated with peptide-MHC binding affinity measurements for at least one of the peptides; and

(b) data associated with peptide-MHC binding stability measurements for at least one of the peptides.

14. The method of any one of claims 1-13, wherein the set of presentation likelihoods are

further identified by at least expression levels of the one or more MHC alleles in the subject, as measured by RNA-seq or mass spectrometry.

15. The method of any one of claims 1-14, wherein the set of presentation likelihoods are further identified by features comprising at least one of:

(a) predicted affinity between a neoantigen in the set of neoantigens and the one or more MHC alleles; and

(b) predicted stability of the neoantigen encoded peptide-MHC complex.

16. The method of any one of claims 1-15, wherein the set of numerical likelihoods are further identified by features comprising at least one of:

(a) the C-terminal sequences flanking the neoantigen encoded peptide sequence within its source protein sequence; and

(b) the N-terminal sequences flanking the neoantigen encoded peptide sequence within its source protein sequence.

17. The method of any one of claims 1-16, wherein selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being presented on the tumor cell surface relative to unselected neoantigens based on the machine-learned presentation model.

18. The method of any one of claims 1-17, wherein selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being capable of inducing a tumor-specific immune response in the subject relative to unselected neoantigens based on the machine-learned presentation model.

19. The method of any one of claims 1-18, wherein selecting the set of selected neoantigens comprises selecting neoantigens that have an increased likelihood of being capable of being presented to naïve T-cells by professional antigen presenting cells (APCs) relative to unselected neoantigens based on the presentation model, optionally wherein the APC is a dendritic cell (DC).

20. The method of any one of claims 1-19, wherein selecting the set of selected neoantigens comprises selecting neoantigens that have a decreased likelihood of being subject to inhibition via central or peripheral tolerance relative to unselected neoantigens based on the machine-learned presentation model.

21. The method of any one of claims 1-20, wherein selecting the set of selected neoantigens comprises selecting neoantigens that have a decreased likelihood of being capable of inducing an autoimmune response to normal tissue in the subject relative to unselected neoantigens based on the machine-learned presentation model.

22. The method of any one of claims 1-21, wherein the one or more tumor cells are selected from the group consisting of: lung cancer, melanoma, breast cancer, ovarian cancer, prostate cancer, kidney cancer, gastric cancer, colon cancer, testicular cancer, head and neck cancer, pancreatic cancer, brain cancer, B-cell lymphoma, acute myelogenous leukemia, chronic myelogenous leukemia, chronic lymphocytic leukemia, and T-cell lymphocytic leukemia, non-small cell lung cancer, and small cell lung cancer.

23. The method of any one of claims 1-22, further comprising generating an output for constructing a personalized cancer vaccine from the set of selected neoantigens.

24. The method of claim 23, wherein the output for the personalized cancer vaccine comprises at least one peptide sequence or at least one nucleotide sequence encoding the set of selected neoantigens.

25. The method of any one of claims 1-24, wherein the machine-learned presentation model is a neural network model.

26. The method of claim 25, wherein the neural network model comprises a single neural network model including a series of nodes arranged in one or more layers, the single neural netowrk model configured to receive numerical vectors encoding the peptide sequences of multiple different MHC alleles.

27. The method of claim 26, wherein the neural network model is trained by updating the parameters of the neural network model.

28. The method of any one of claims 25-27, wherein the machine-learned presentation model is a deep learning model that includes one or more layers of nodes.

29. The method of any one of claims 1-28,

wherein the training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the at least one MHC allele bound to the peptides of the sample and a set of positions of the amino acids in the at least one MHC allele, do not include a peptide sequence of a MHC allele of the subject that is input into the machine-learned presentation model to generate the

set of presentation likelihoods for the set of neoantigens.

30. The method of any one of claims 1-29, wherein the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set belongs to a gene family to which the one or more MHC alleles of the subject belongs.

31. The method of any one of claims 1-30, wherein the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set comprises one MHC allele.

32. The method of any one of claims 1-30, wherein the at least one MHC allele bound to the peptides of each sample of the plurality of samples of the training data set comprises more than one MHC allele.

33. The method of any one of claims 1-32, wherein the one or more MHC alleles are class I MHC alleles.

34. A computer system comprising:

a computer processor;

a memory storing computer program instructions that when executed by the computer processor cause the computer processor to:

obtain at least one of exome, transcriptome, or whole genome nucleotide sequencing data from the tumor cells and normal cells of the subject, wherein the nucleotide sequencing data is used to obtain data representing peptide sequences of each of a set of neoantigens identified by comparing the nucleotide sequencing data from the tumor cells and the nucleotide sequencing data from the normal cells, wherein the peptide sequence of each neoantigen comprises at least one alteration that makes it distinct from the corresponding wild-type peptide sequence identified from the normal cells of the subject;

encodethe peptide sequences of each of the neoantigens into a corresponding numerical vector, each numerical vector including information regarding a plurality of amino acids that make up the peptide sequence and a set of positions of the amino acids in the peptide sequence;

obtain at least one of exome, transcriptome or whole genome nucleotide sequencing data from each of the one or more MHC alleles of the subject, wherein the

nucleotide sequencing data is used to obtain data representing a peptide sequence of each of the one or more MHC alleles of the subject;

encode the peptide sequences of each of the one or more MHC alleles of the subject into a corresponding numerical vector, each numerical vector including information regarding a plurality of amino acids that make up the peptide sequence and a set of positions of the amino acids in the peptide sequence;

input the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors encoding the peptide sequences of each of the one or more MHC alleles, using a computer processor, into a machine-learned presentation model to generate a set of presentation likelihoods for the set of neoantigens, each presentation likelihood in the set representing the likelihood that a corresponding neoantigen is presented by the one or more MHC alleles on the surface of the tumor cells of the subject, the machine-learned presentation model comprising:

a plurality of parameters identified at least based on a training data set comprising:

for each sample in a plurality of samples, a label obtained by mass spectrometry measuring presence of peptides bound to at least one MHC allele in a set of MHC alleles identified as present in the sample;

for each of the samples, training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the peptides and a set of positions of the amino acids in the peptides; and

for each of the samples, training peptide sequences encoded as numerical vectors including information regarding a plurality of amino acids that make up the at least one MHC allele bound to the peptides of the sample and a set of positions of the amino acids in the at least one MHC allele;

a function representing a relation between the numerical vectors encoding the peptide sequences of each of the neoantigens and the numerical vectors

encoding the peptide sequences of each of the one or more MHC alleles received as input, and the presentation likelihood generated as output based on the numerical vectors and the parameters;

select a subset of the set of neoantigens based on the set of presentation likelihoods to generate a set of selected neoantigens; and

return the set of selected neoantigens.

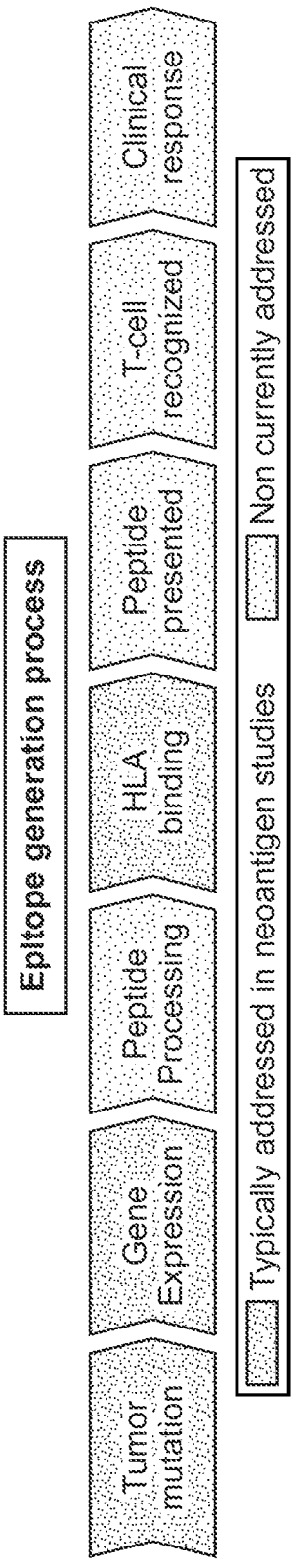
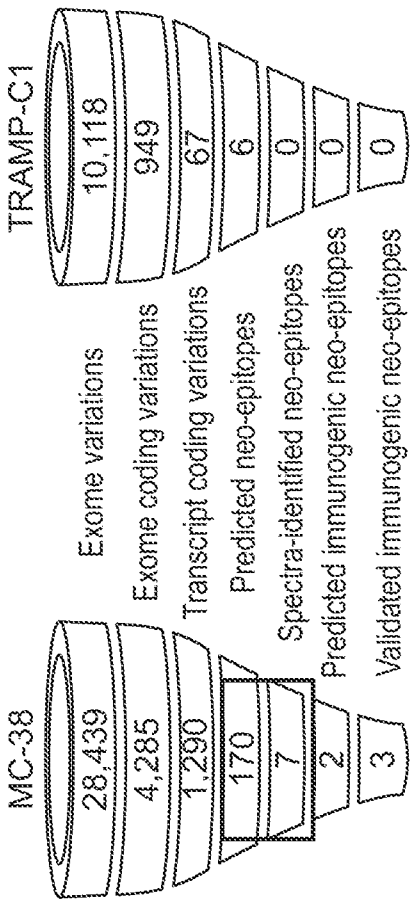

Epitope generation process
Tumor mutation
Gene Expression
Peptide Processing
HLA binding
Peptide presented
T-cell recognized
Clinical response
Typically addressed in neoantigen studies
Non currently addressed
Common prediction pipeline steps
Exome
RNASeq
NetMHC

FIG. 1A



Prediction	NetMHCcon IC50≤500nM	MS-based identification HLAp	Not detected	
	Binders	1,579	47,777	3%
	Non-binders	153	1,140,967	

Prediction	NetMHCcon IC50≤50nM	MS-based identification HLAp	Not detected
	Binders	1,133	14,059
	Non-binders	599	1,174,685

FIG. 1B

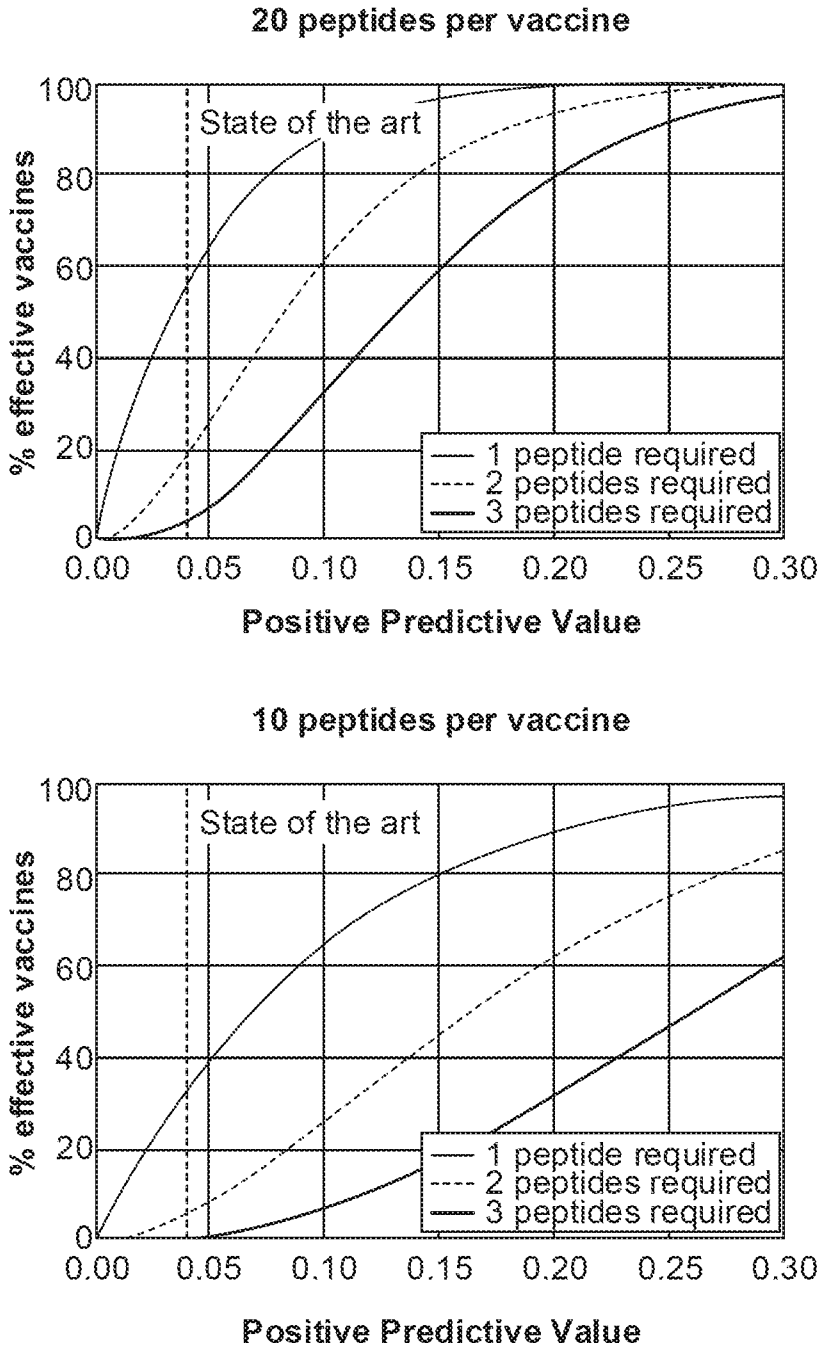

20 peptides per vaccine
State of the art
% effective vaccines
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Positive Predictive Value
1 peptide required
2 peptides required
3 peptides required
10 peptides per vaccine
State of the art
% effective vaccines
Positive Predictive Value
1 peptide required
2 peptides required
3 peptides required

FIG. 1C

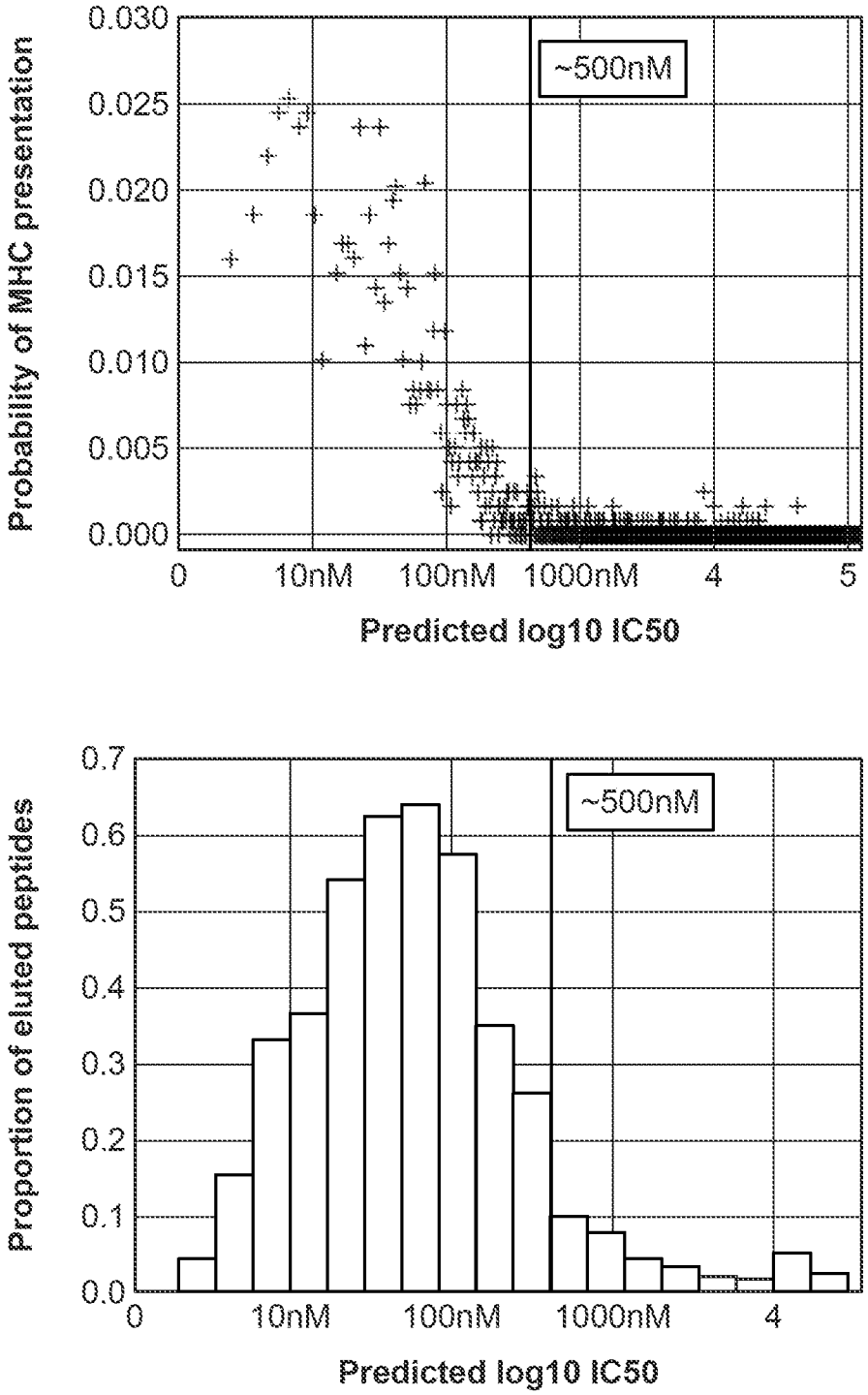

Probability of MHC presentation
0.030
0.025
0.020
0.015
0.010
0.005
0.000
~500nM
0
10nM
100nM
1000nM
4
5
Predicted log10 IC50
Proportion of eluted peptides
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
~500nM
0
10nM
100nM
1000nM
4
Predicted log10 IC50

FIG. 1D

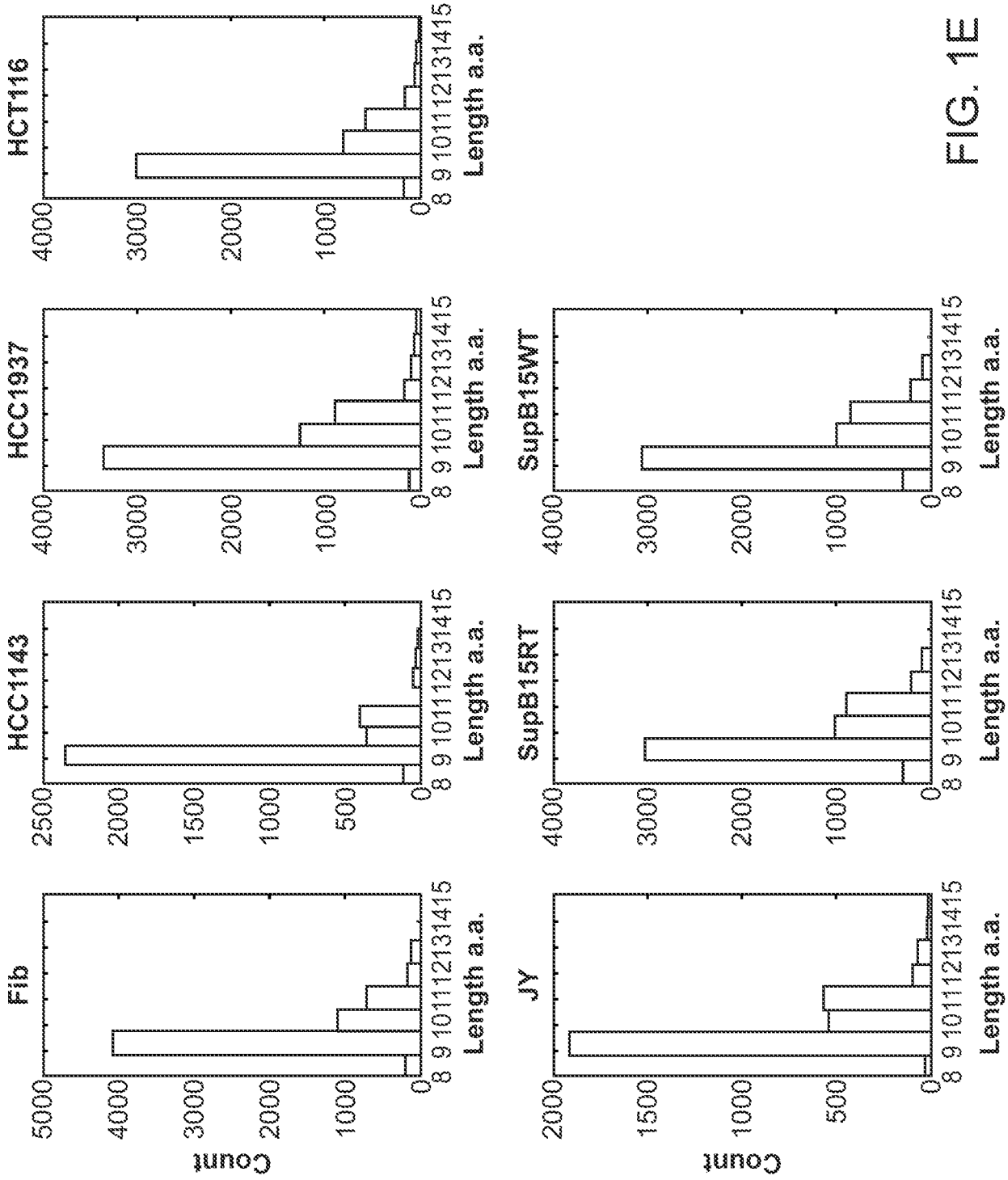

Fib
HCC1143
HCC1937
HCT116
JY
SupB15RT
SupB15WT
Count
Length a.a.
8 9 10 11 12 13 14 15
0
500
1000
1500
2000
2500
3000
4000
5000

FIG. 1E

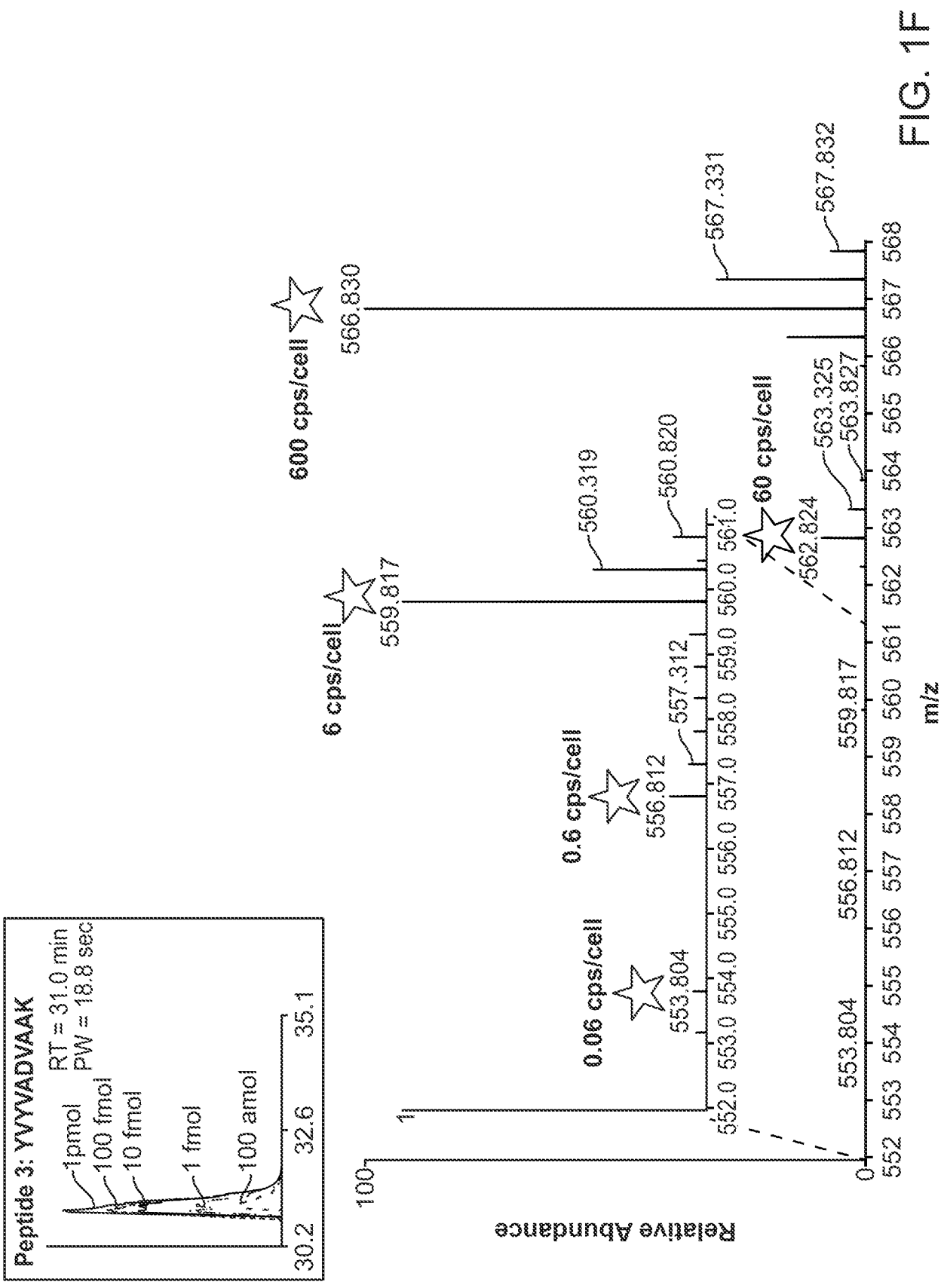

Peptide 3: YVYVADVAAK
RT = 31.0 min
PW = 18.8 sec
1pmol
100 fmol
10 fmol
1 fmol
100 amol
30.2
32.6
35.1
Relative Abundance
m/z
600 cps/cell
6 cps/cell
0.6 cps/cell
0.06 cps/cell
60 cps/cell
566.830
559.817
560.319
560.820
557.312
556.812
553.804
562.824
563.325
563.827
567.331
567.832

FIG. 1F

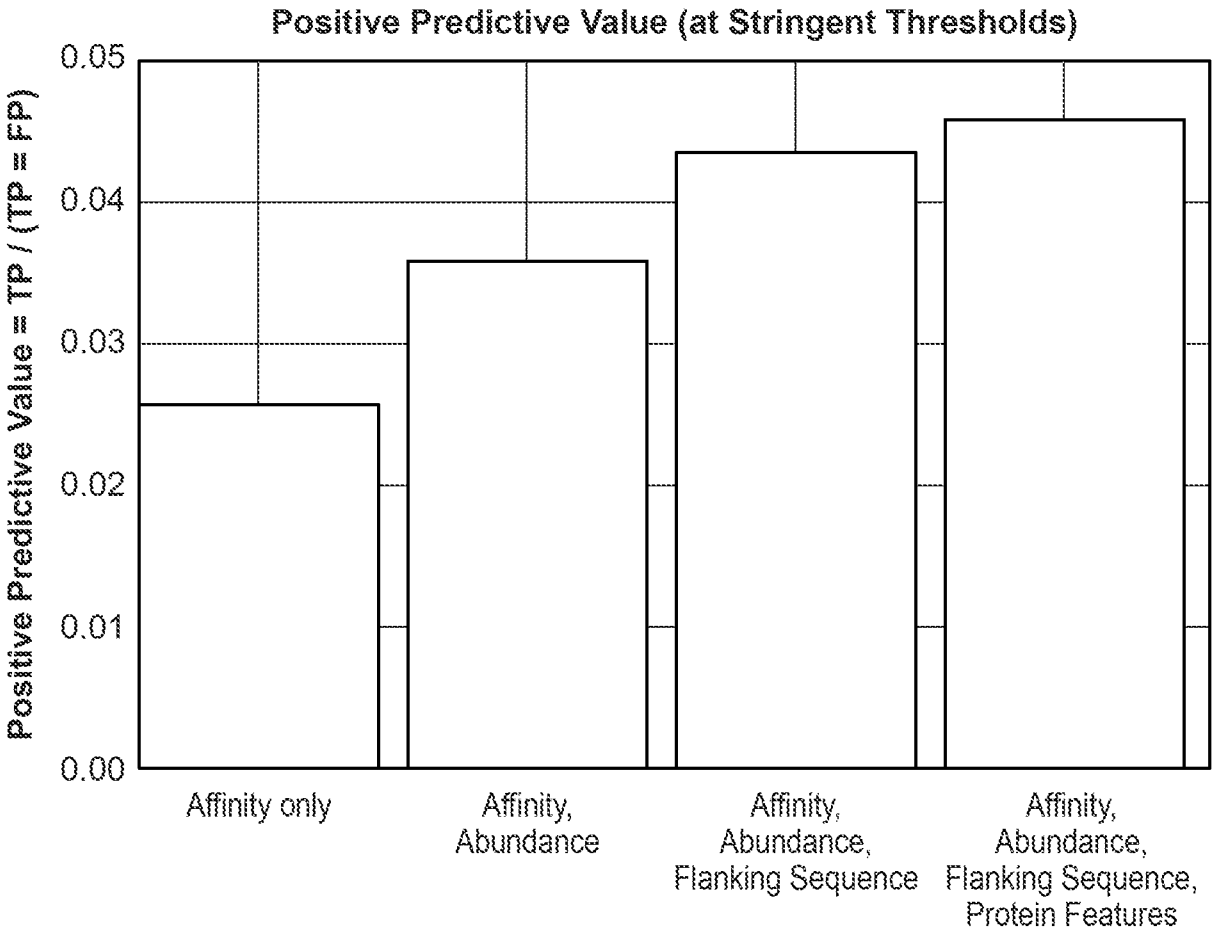

Positive Predictive Value (at Stringent Thresholds)
Positive Predictive Value = TP / (TP + FP)
0.05
0.04
0.03
0.02
0.01
0.00
Affinity only
Affinity, Abundance
Affinity, Abundance, Flanking Sequence
Affinity, Abundance, Flanking Sequence, Protein Features

FIG. 1G

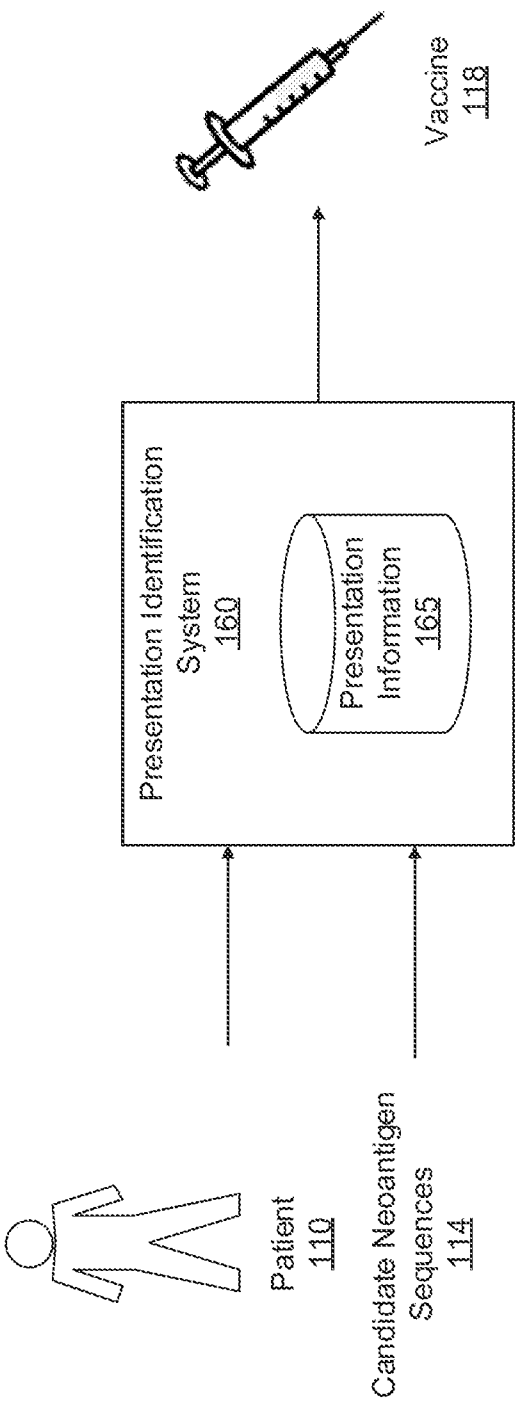

Patient
110
Candidate Neoantigen
Sequences
114
Presentation Identification
System
160
Presentation
Information
165
Vaccine
118

FIG. 2A

Single-Allele Cell Lines
Peptide Isolation
HLA-DRB1*12:01
YEMFNDKSFQRAPDDKMF

FIG. 2B

Multiple-Allele Cell Lines
Peptide Isolation
HLA-A*01:01
HLA-A*02:01
HLA-B*07:02
HLA-B*08:01
HLA-DRB1*10:01
HLA-DRB1:11:01
YEMFNDKSF
HROEIFSHDFJ
FJIEJFOESS
NEIOREIREI
JFKSIFEMMSJDSSUIFLKSJFIEIFJ
KNFLENFIESOFI

FIG. 2C

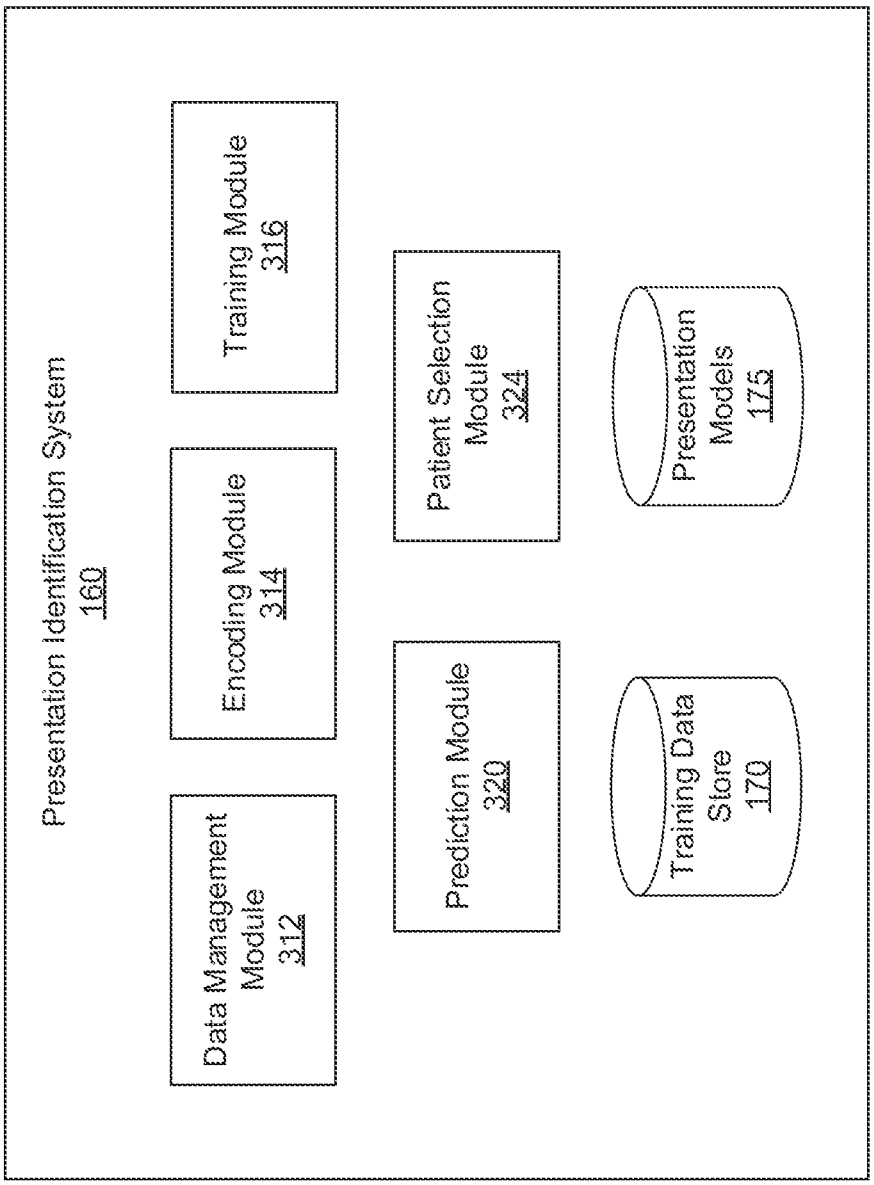

Presentation Identification System
160
Data Management Module
312
Encoding Module
314
Training Module
316
Prediction Module
320
Patient Selection Module
324
Training Data Store
170
Presentation Models
175

FIG. 3

Training Data
170A

Allele-Dependent (x^i) | Allele-Independent (w^i)

Peptide Sequence (p^i)	Affinity (b^i-nM)	Stability (s^i-h)	Allele (a^i)	C-Flanking Sequence (c^i)	mRNA Q. (m^i-*TPM*)	Label (y^i)
QCEIOWAREFLKEIGJ	1000	1	HLA-DRB3:01:01	FJELFISBOSJFIE	10^2	Not Presented
FIEUHFWI	1500	15	HLA-C*01:03	FEGRKUOOI	10^{-3}	Presented
FEWRHRJTRUJR	650	20	HLA-C*01:03	PJFIOEJOIJGEIO	10^1	Presented
QIEJOEIJE	500	1	HLA-B*07:02	PJFIOEJOIJGEIO	1	Presented
	600	14	HLA-C*01:03			
	1200	7	HLA-A*01:01			

FIG. 4

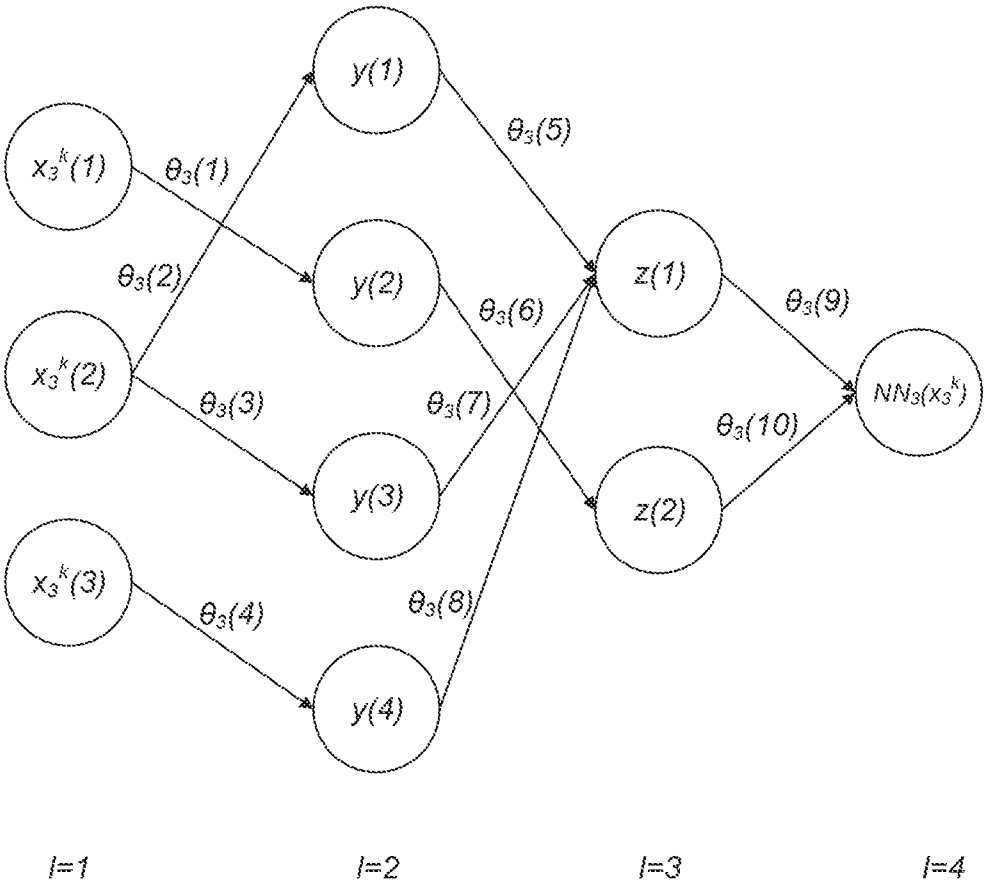

$x_3^k(1)$
$x_3^k(2)$
$x_3^k(3)$
$\theta_3(1)$
$\theta_3(2)$
$\theta_3(3)$
$\theta_3(4)$
y(1)
y(2)
y(3)
y(4)
$\theta_3(5)$
$\theta_3(6)$
$\theta_3(7)$
$\theta_3(8)$
z(1)
z(2)
$\theta_3(9)$
$\theta_3(10)$
$NN_3(x_3^k)$
l=1
l=2
l=3
l=4

FIG. 5

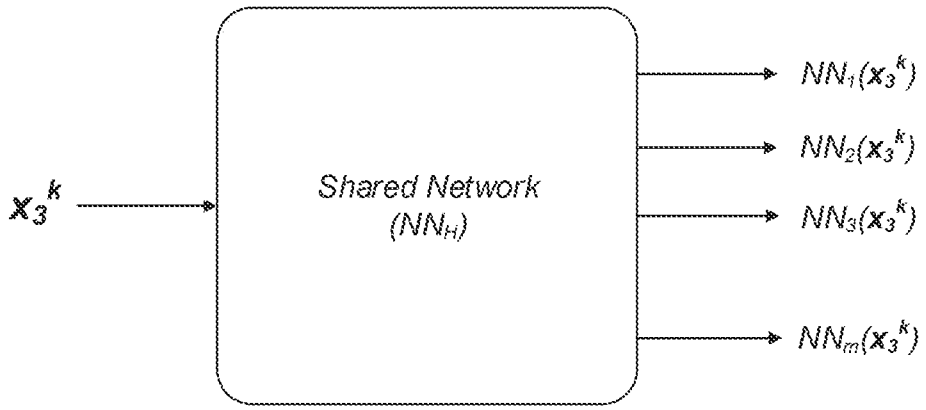


FIG. 6

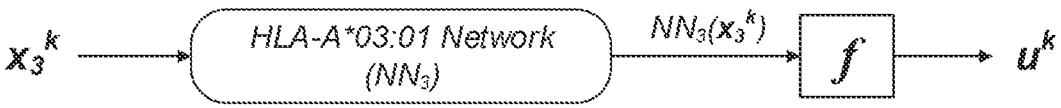


FIG. 7

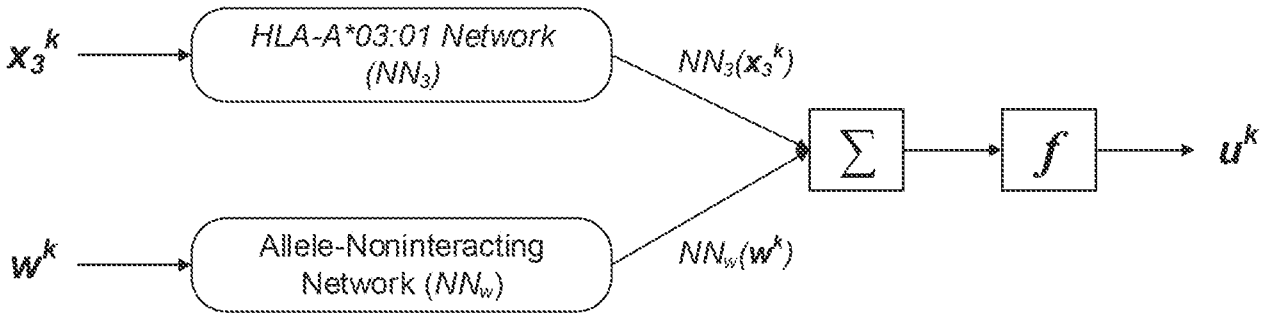

x_3^k
HLA-A*03:01 Network (NN_3)
$NN_3(x_3^k)$
w^k
Allele-Noninteracting Network (NN_w)
$NN_w(w^k)$
$\sum$
f
u^k

FIG. 8

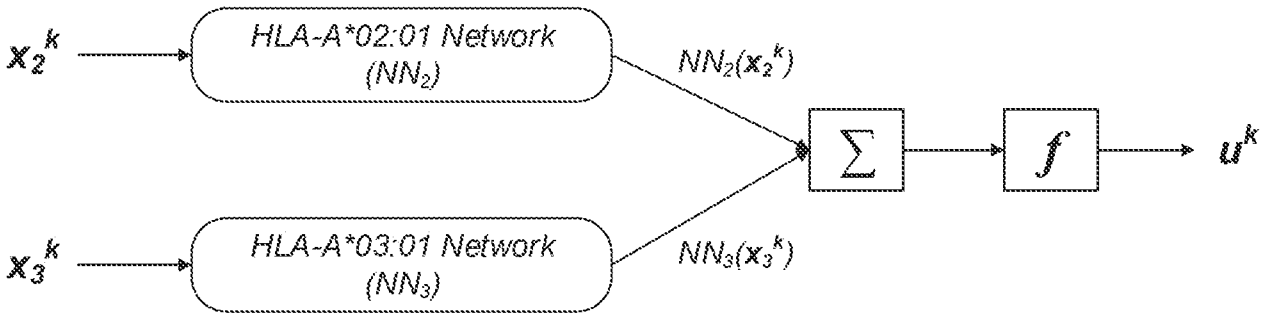

x_2^k
HLA-A*02:01 Network
(NN_2)
$NN_2(x_2^k)$
x_3^k
HLA-A*03:01 Network
(NN_3)
$NN_3(x_3^k)$
$\sum$
f
u^k

FIG. 9

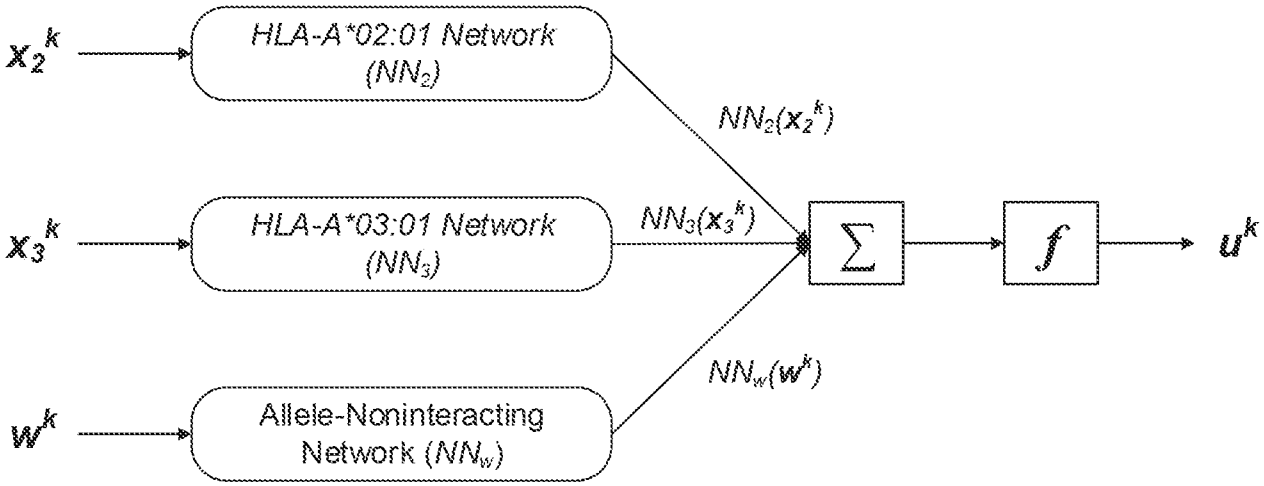

x_2^k
HLA-A*02:01 Network (NN_2)
$NN_2(x_2^k)$
x_3^k
HLA-A*03:01 Network (NN_3)
$NN_3(x_3^k)$
$\sum$
f
u^k
w^k
Allele-Noninteracting Network (NN_w)
$NN_w(w^k)$

FIG. 10

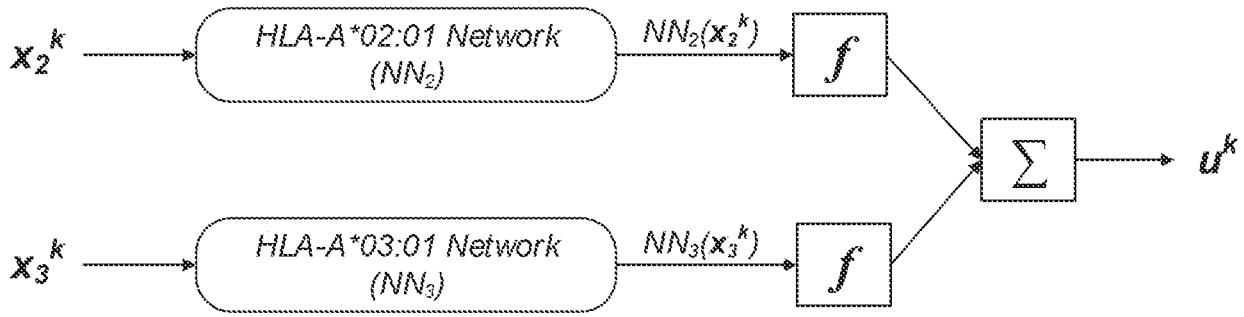

x_2^k
HLA-A*02:01 Network
(NN_2)
$NN_2(x_2^k)$
f
x_3^k
HLA-A*03:01 Network
(NN_3)
$NN_3(x_3^k)$
f
$\sum$
u^k

FIG. 11

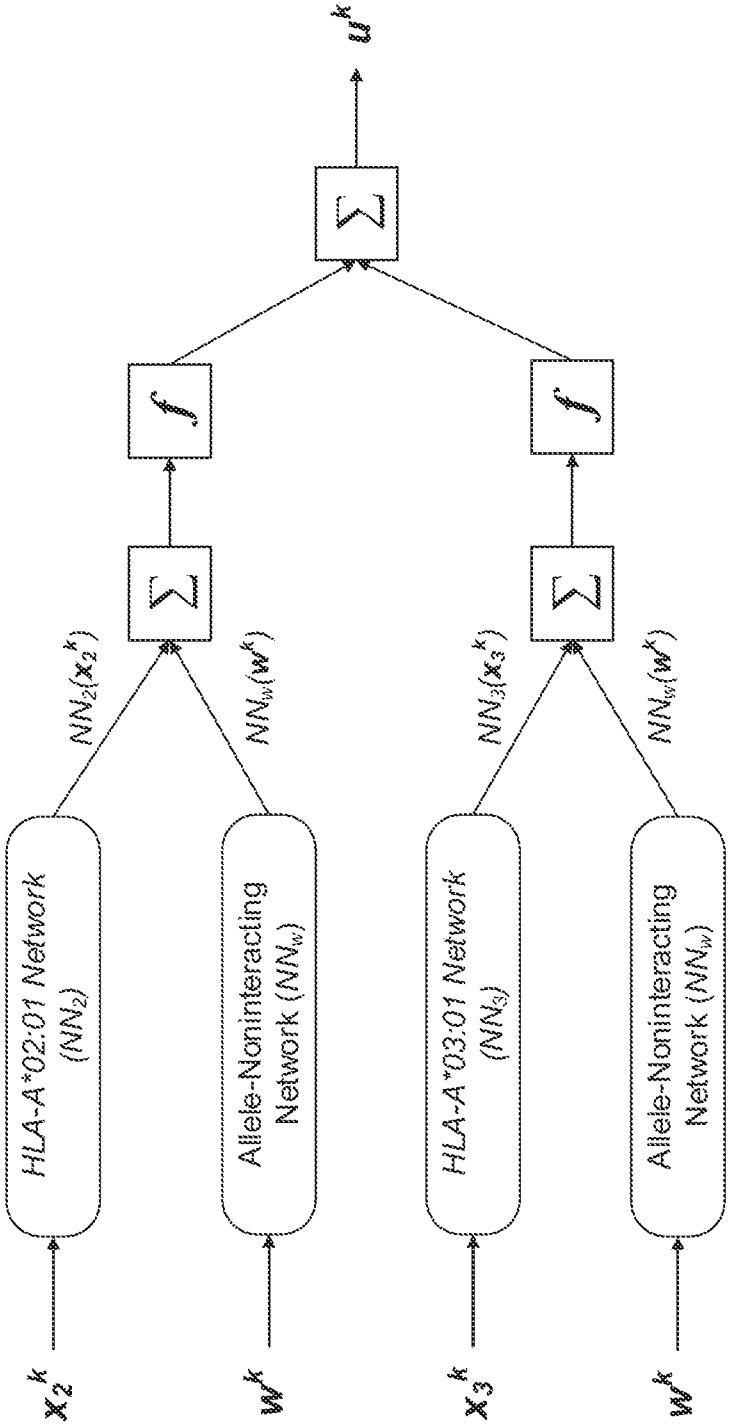

u^k
$\sum$
f
f
$\sum$
$\sum$
$NN_2(x_2^k)$
$NN_w(w^k)$
$NN_3(x_3^k)$
$NN_w(w^k)$
HLA-A*02:01 Network (NN_2)
Allele-Noninteracting Network (NN_w)
HLA-A*03:01 Network (NN_3)
Allele-Noninteracting Network (NN_w)
x_2^k
w^k
x_3^k
w^k

FIG. 12

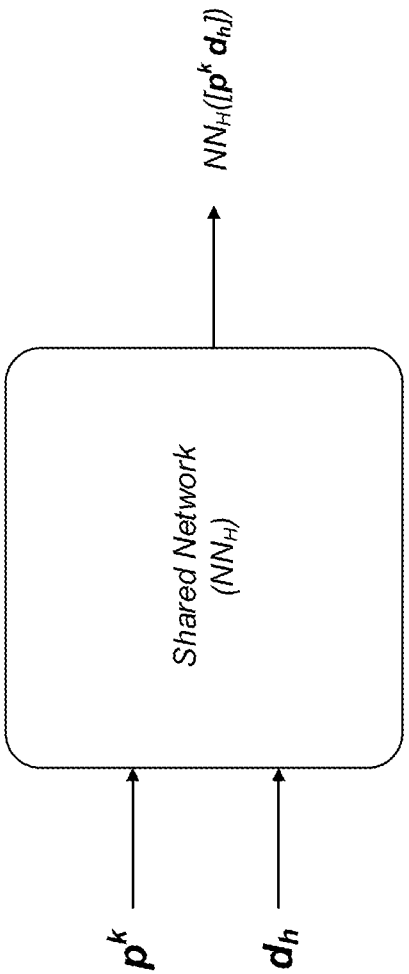


FIG. 13

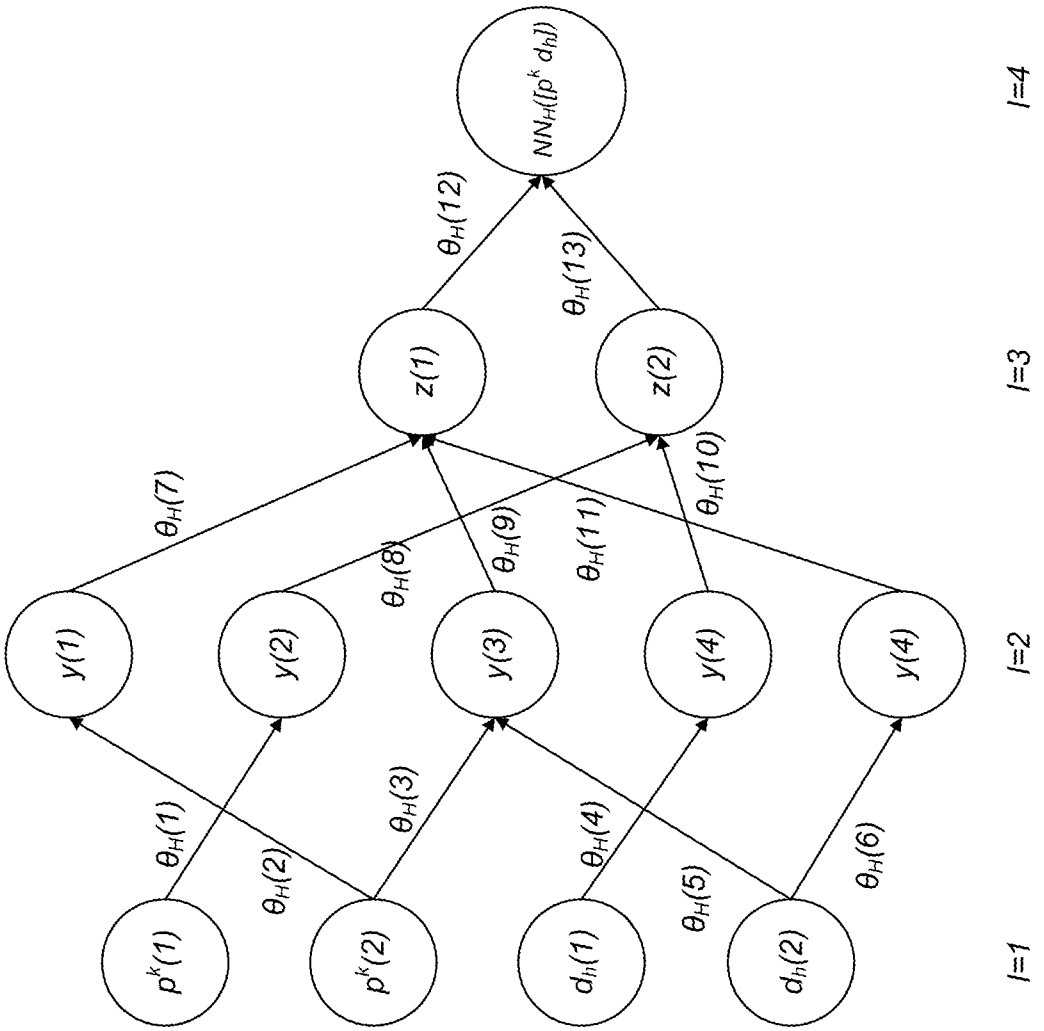

$p^k(1)$
$p^k(2)$
$d_{hI}(1)$
$d_{hI}(2)$
$y(1)$
$y(2)$
$y(3)$
$y(4)$
$y(4)$
$z(1)$
$z(2)$
$NN_H([p^k\ d_{hJ}])$
$\theta_H(1)$
$\theta_H(2)$
$\theta_H(3)$
$\theta_H(4)$
$\theta_H(5)$
$\theta_H(6)$
$\theta_H(7)$
$\theta_H(8)$
$\theta_H(9)$
$\theta_H(10)$
$\theta_H(11)$
$\theta_H(12)$
$\theta_H(13)$
$l=1$
$l=2$
$l=3$
$l=4$

FIG. 14

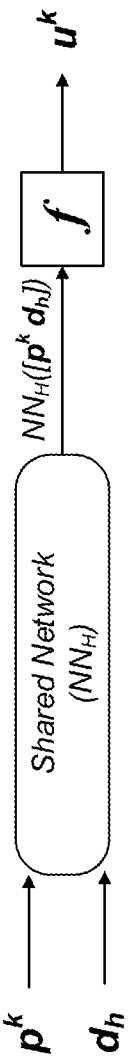


FIG. 15

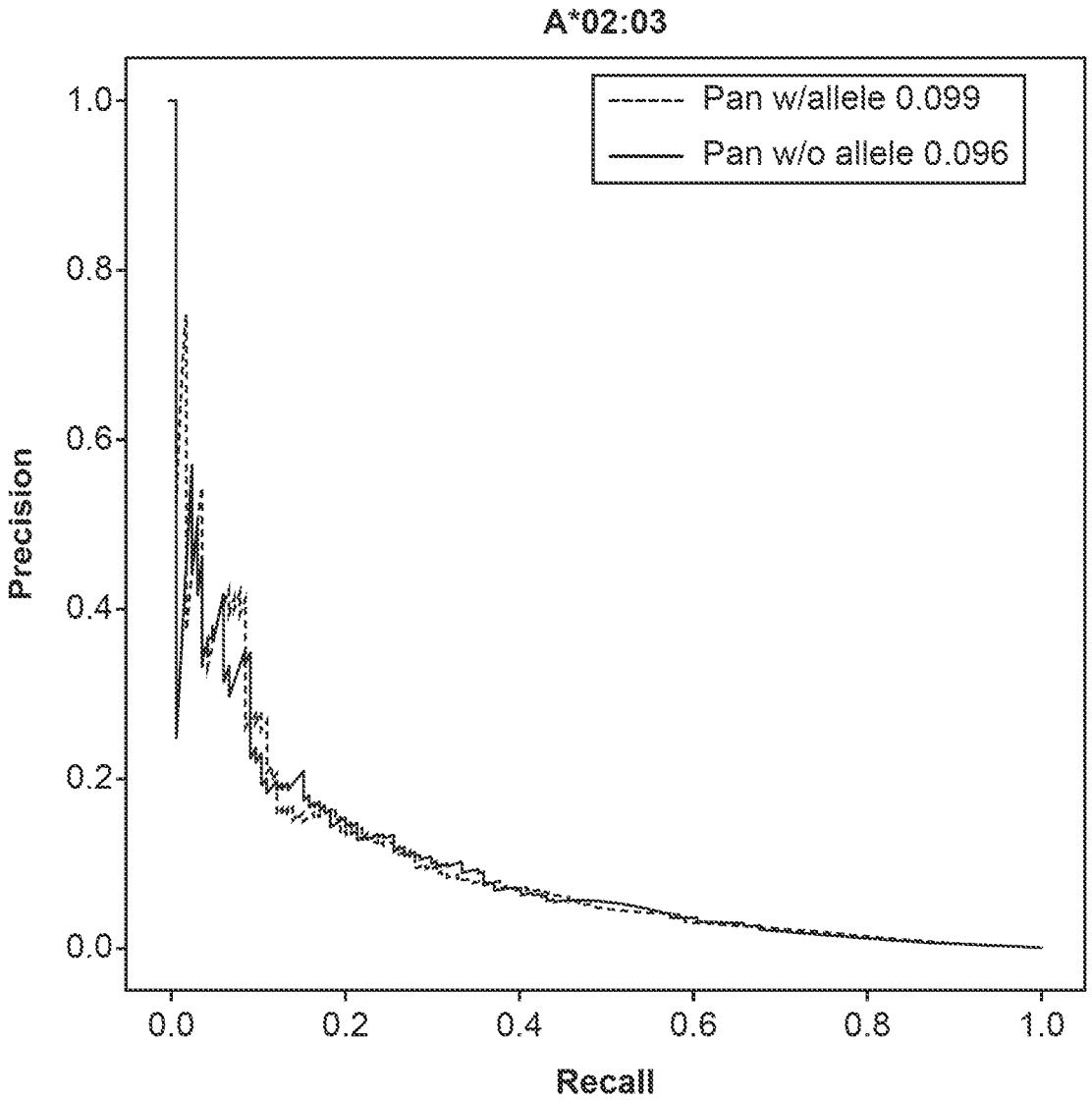

A*02:03
Pan w/allele 0.099
Pan w/o allele 0.096
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

FIG. 16

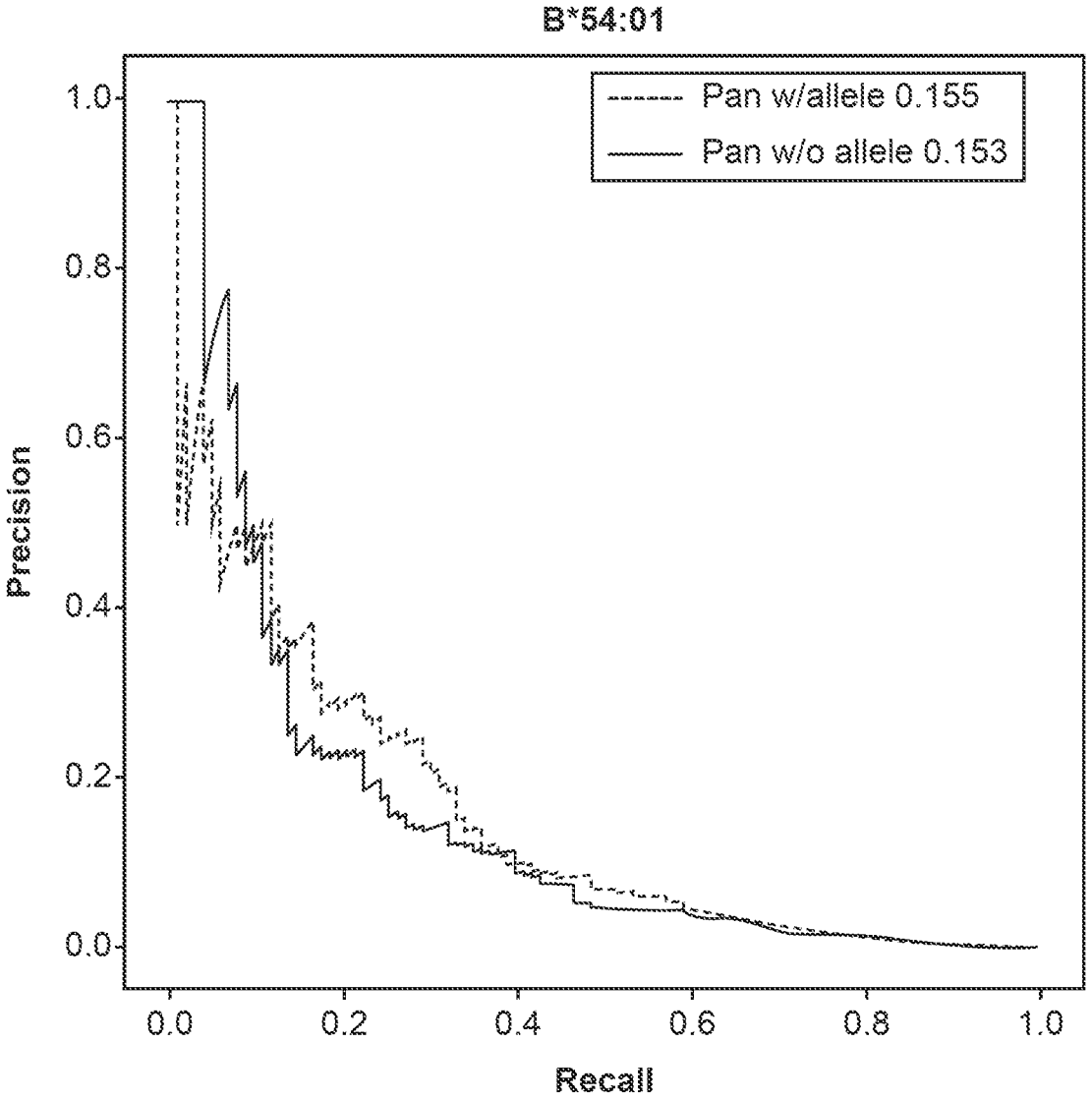

B*54:01
Pan w/allele 0.155
Pan w/o allele 0.153
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 17

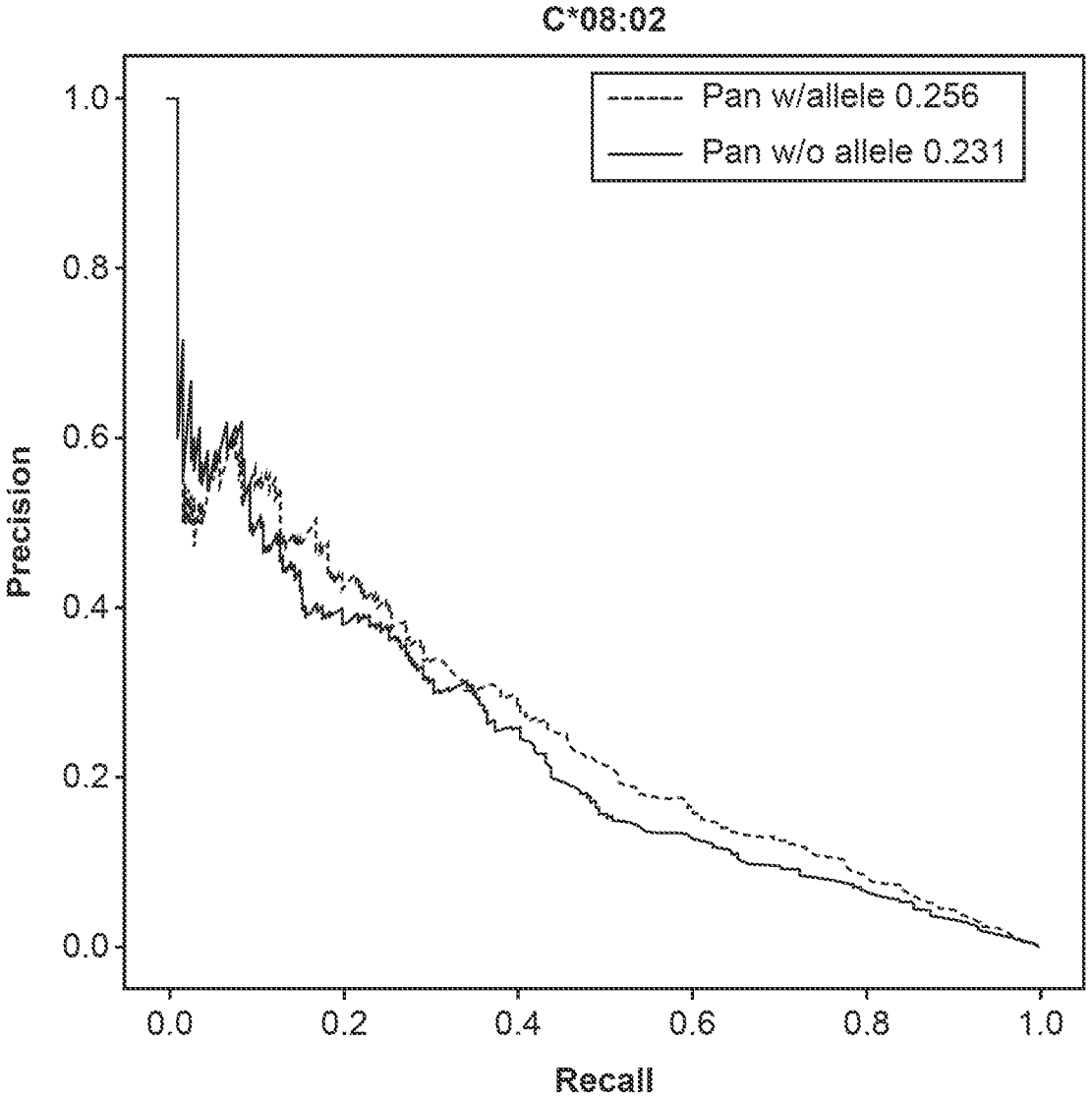

C*08:02
Pan w/allele 0.256
Pan w/o allele 0.231
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 18

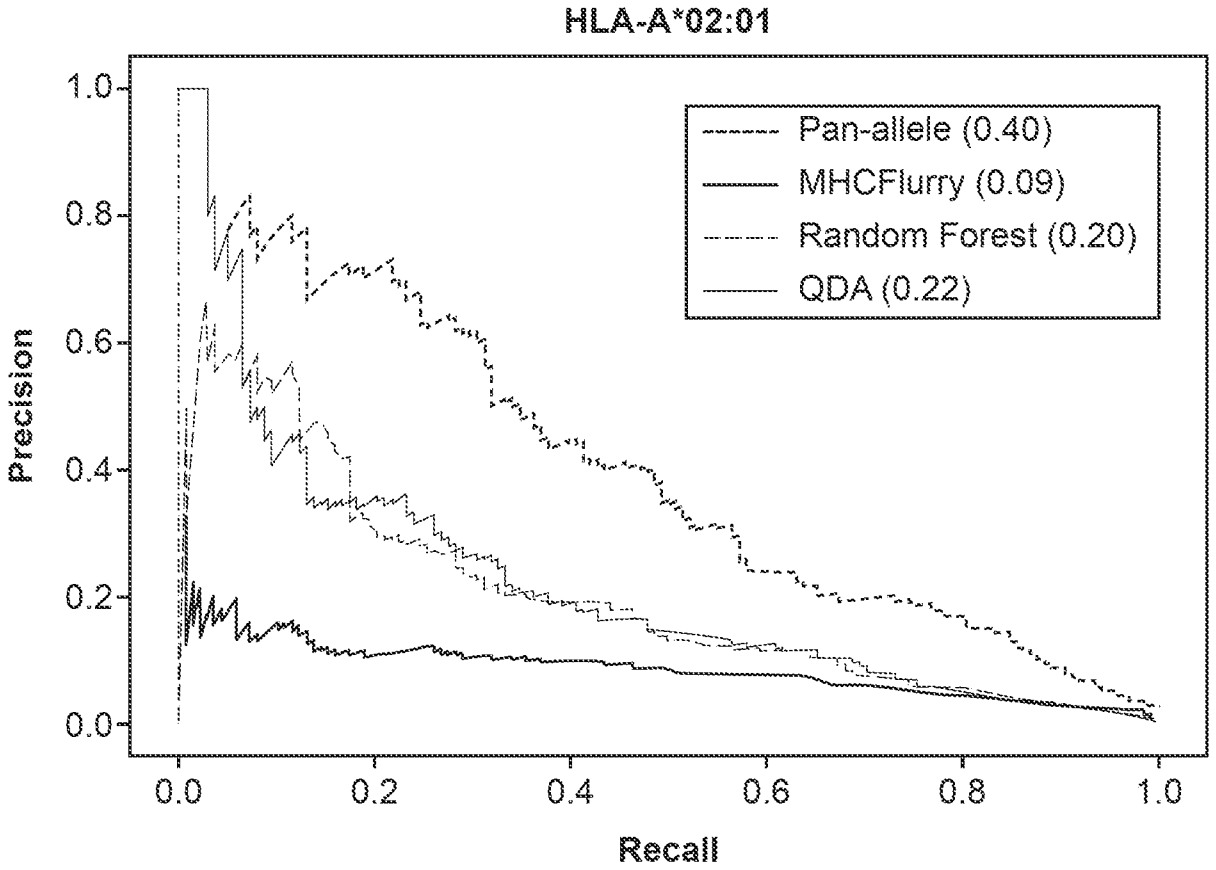

HLA-A*02:01
Pan-allele (0.40)
MHCFlurry (0.09)
Random Forest (0.20)
QDA (0.22)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 19

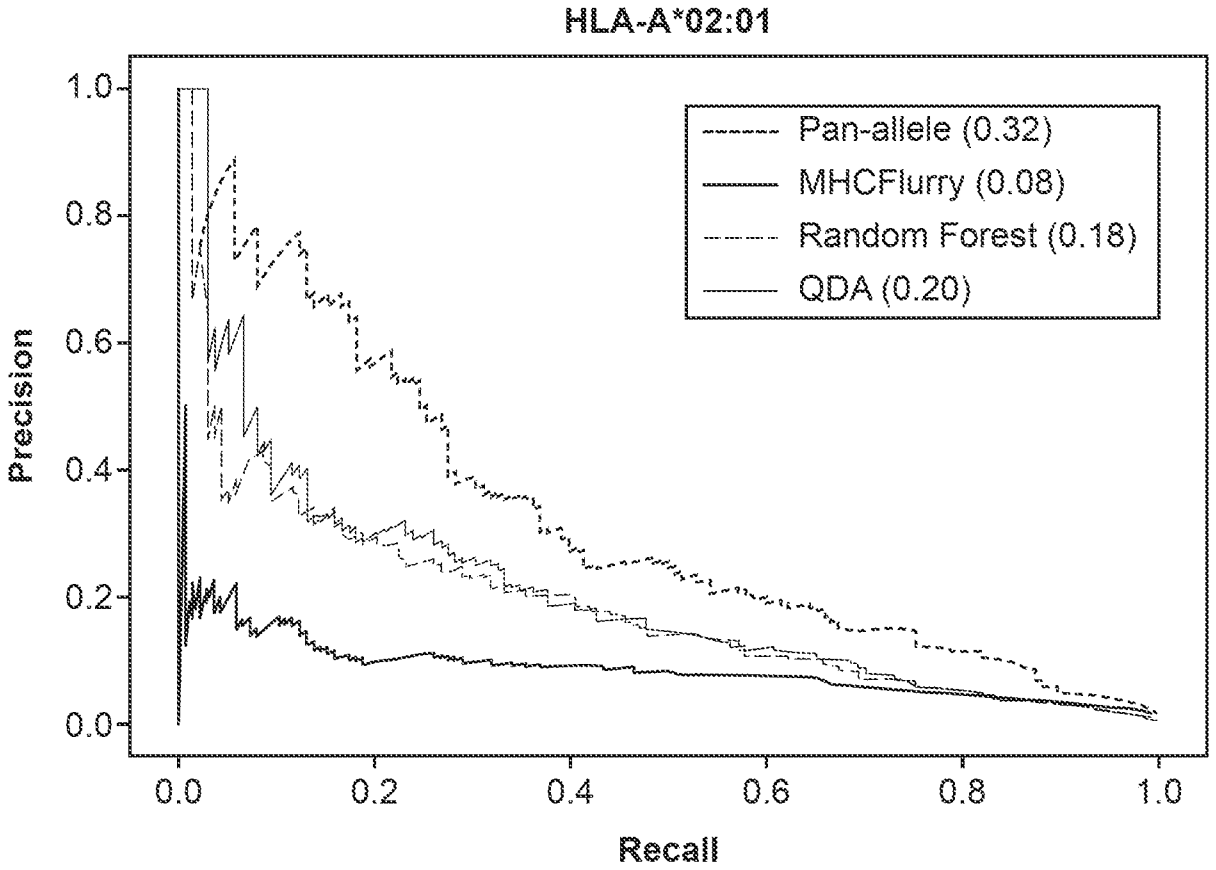

HLA-A*02:01
Pan-allele (0.32)
MHCFlurry (0.08)
Random Forest (0.18)
QDA (0.20)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 20

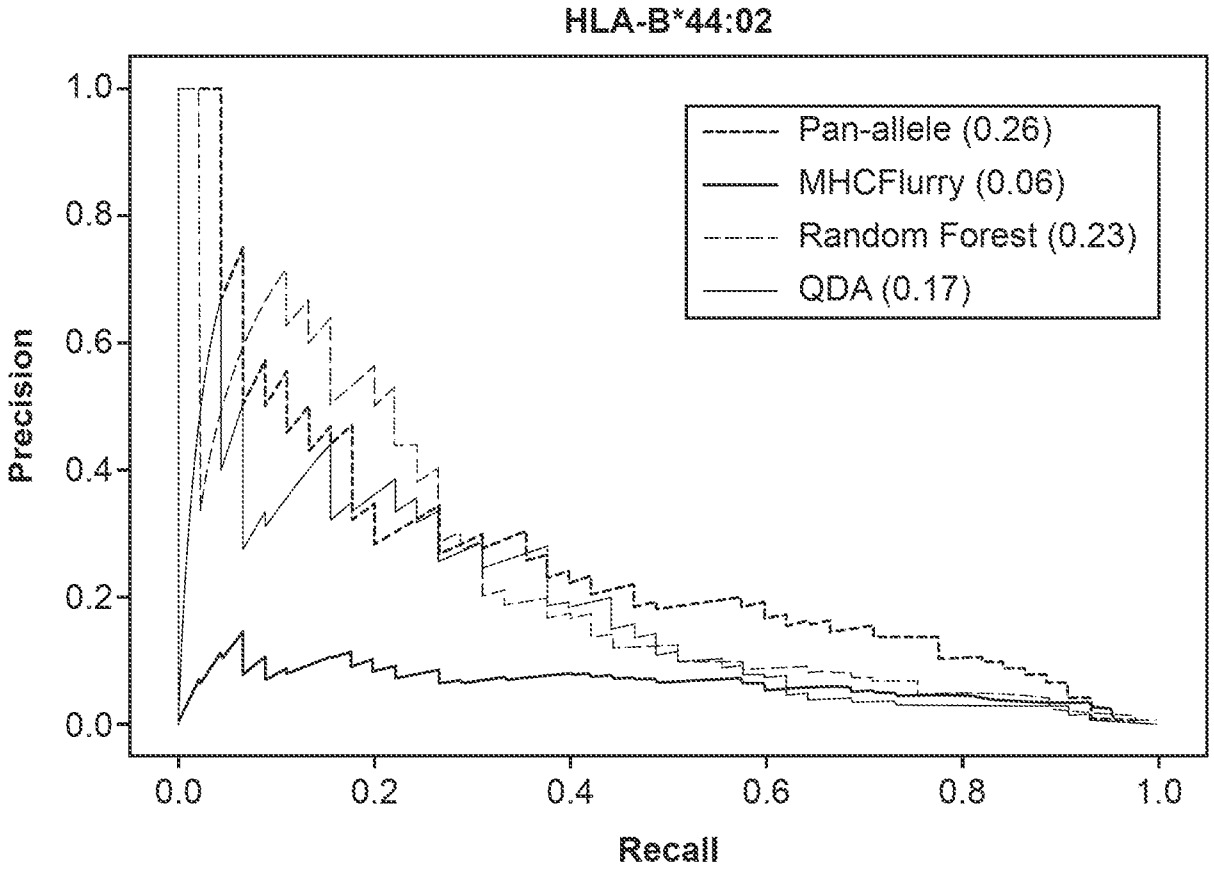

HLA-B*44:02
Pan-allele (0.26)
MHCFlurry (0.06)
Random Forest (0.23)
QDA (0.17)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 21

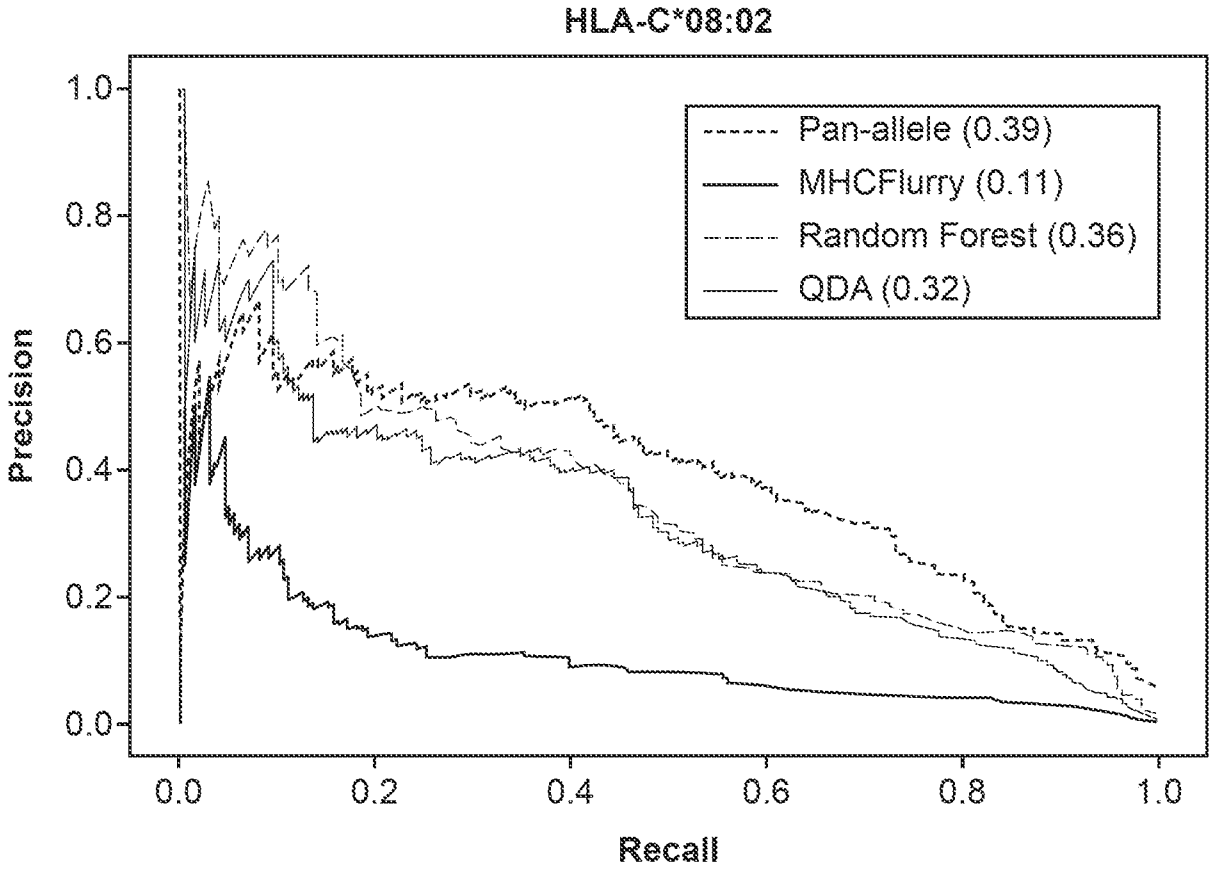

HLA-C*08:02
Pan-allele (0.39)
MHCFlurry (0.11)
Random Forest (0.36)
QDA (0.32)
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 22

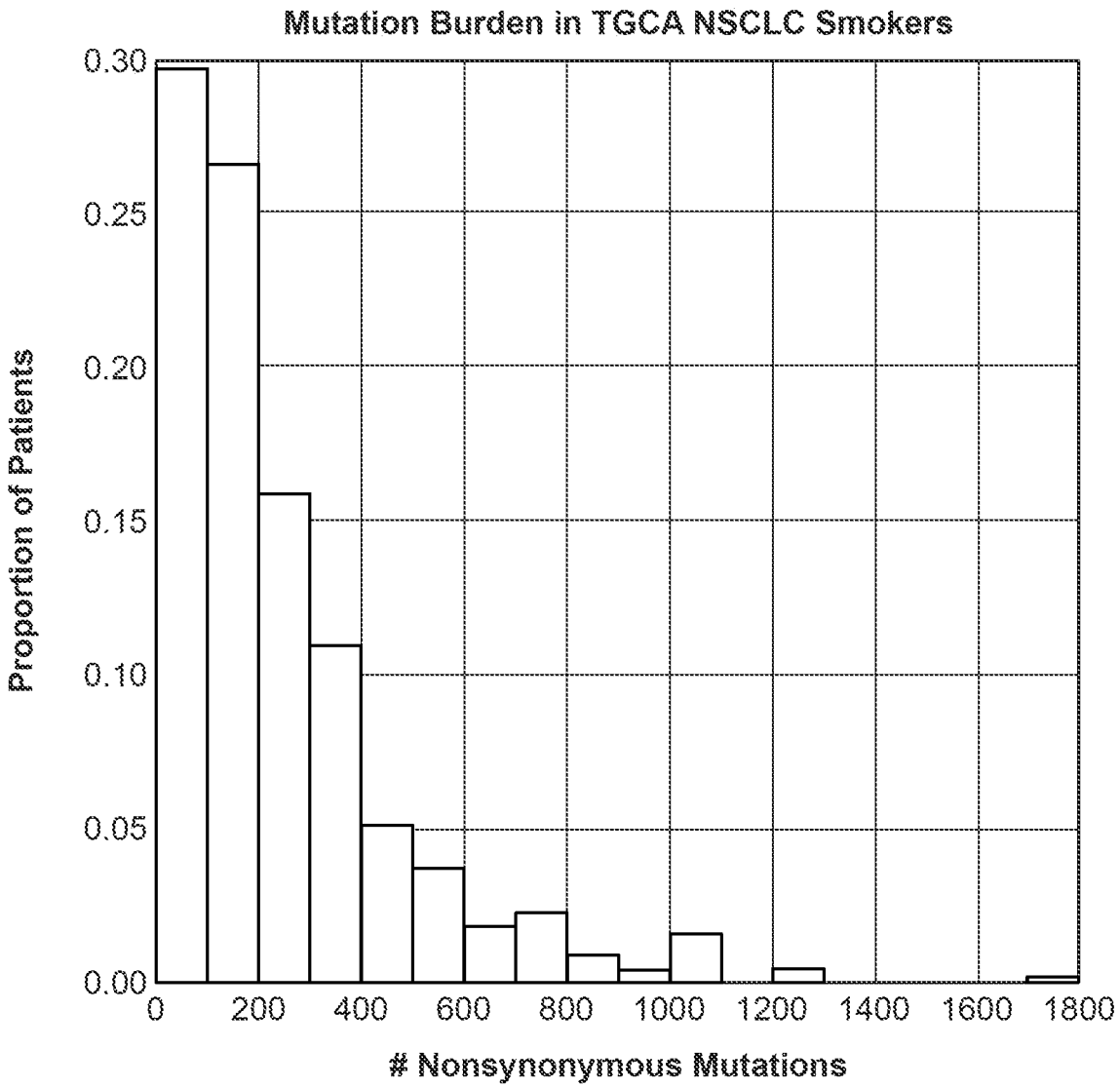

Mutation Burden in TGCA NSCLC Smokers
Proportion of Patients
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
200
400
600
800
1000
1200
1400
1600
1800
Nonsynonymous Mutations

FIG. 23A

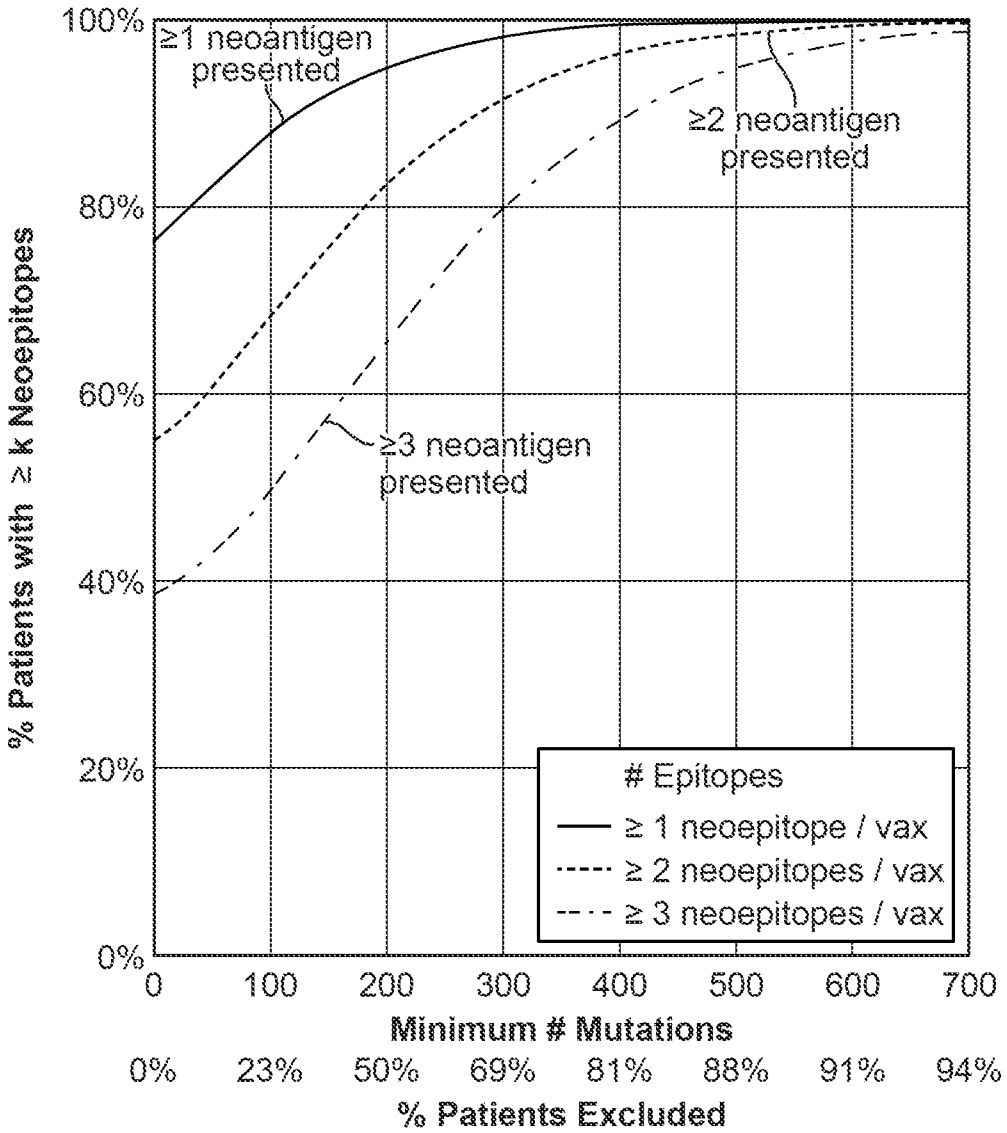

100%
80%
60%
40%
20%
0%
% Patients with ≥ k Neoepitopes
≥1 neoantigen presented
≥2 neoantigen presented
≥3 neoantigen presented
Epitopes
≥ 1 neoepitope / vax
≥ 2 neoepitopes / vax
≥ 3 neoepitopes / vax
0
100
200
300
400
500
600
700
Minimum # Mutations
0%
23%
50%
69%
81%
88%
91%
94%
% Patients Excluded

FIG. 23B

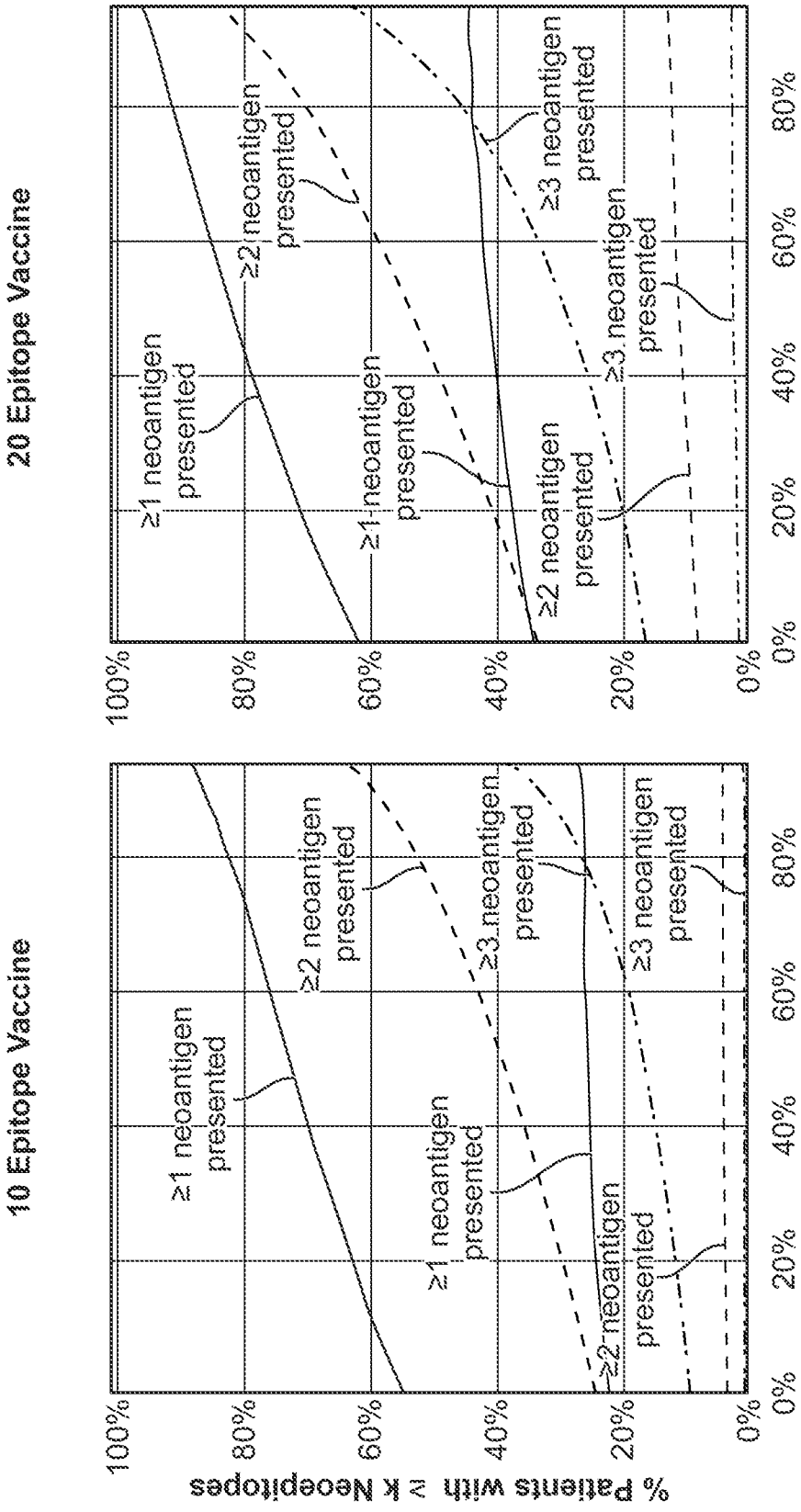

10 Epitope Vaccine
20 Epitope Vaccine
% Patients with ≥ k Neoepitopes
% Patients Excluded
100%
80%
60%
40%
20%
0%
≥1 neoantigen presented
≥2 neoantigen presented
≥3 neoantigen presented

FIG. 23C

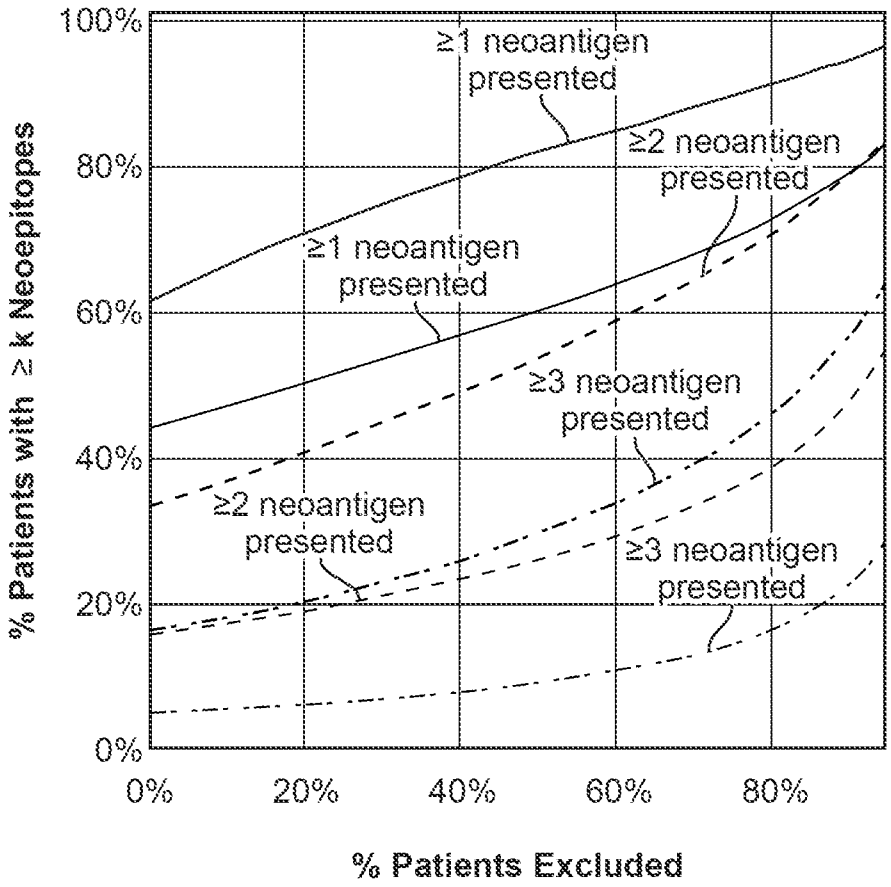

100%
80%
60%
40%
20%
0%
0%
20%
40%
60%
80%
% Patients with ≥ k Neoepitopes
% Patients Excluded
≥1 neoantigen presented
≥2 neoantigen presented
≥1 neoantigen presented
≥3 neoantigen presented
≥2 neoantigen presented
≥3 neoantigen presented

FIG. 23D

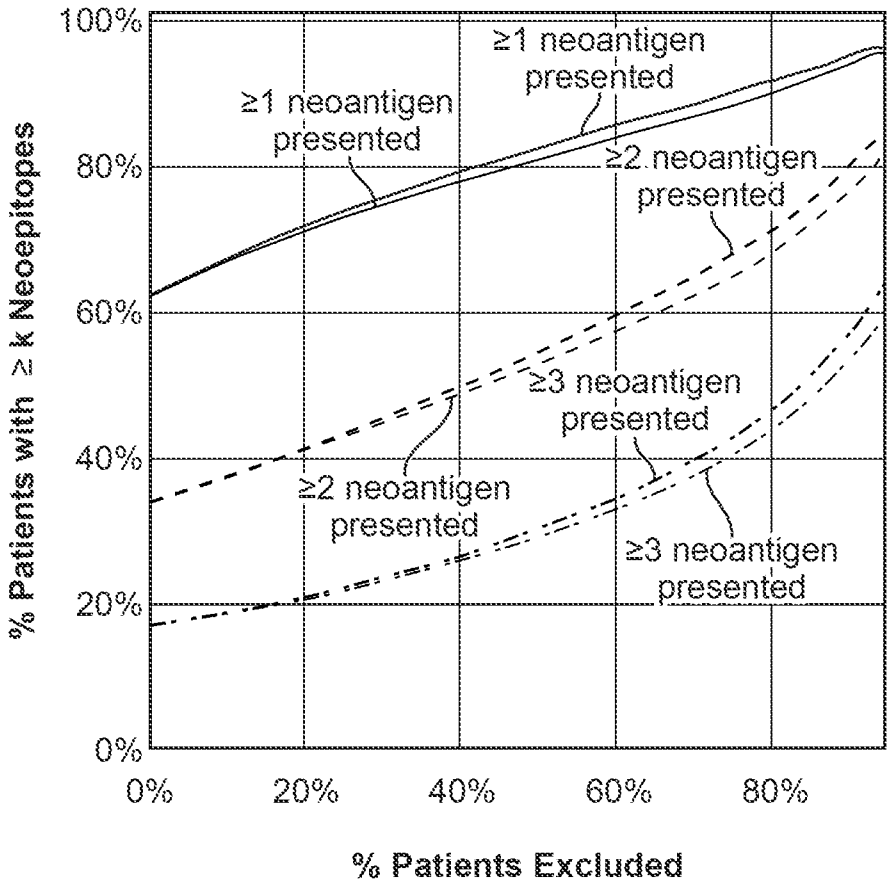

100%
80%
60%
40%
20%
0%
0%
20%
40%
60%
80%
% Patients with ≥ k Neoepitopes
% Patients Excluded
≥1 neoantigen presented
≥1 neoantigen presented
≥2 neoantigen presented
≥2 neoantigen presented
≥3 neoantigen presented
≥3 neoantigen presented

FIG. 23E

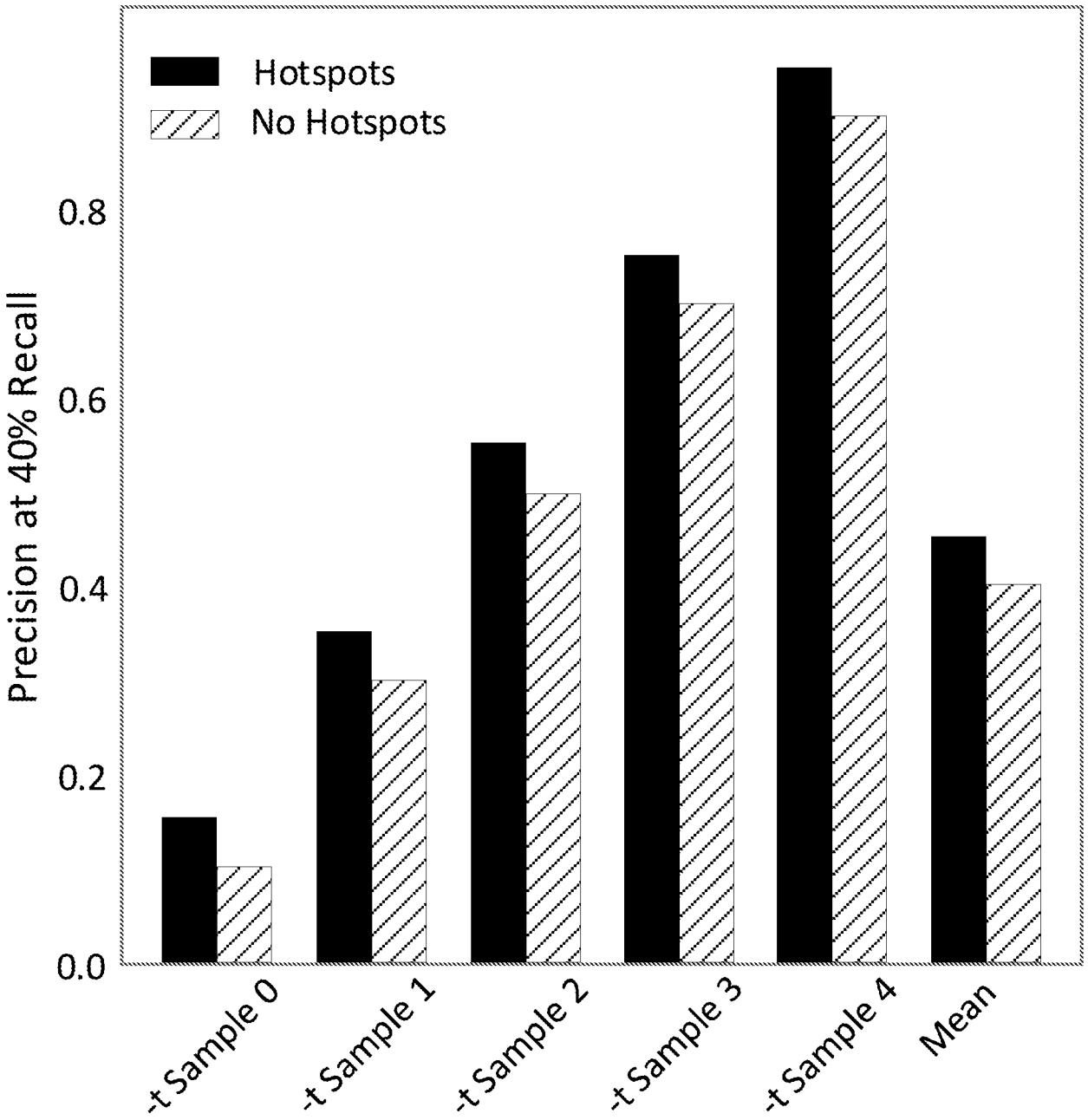

Hotspots
No Hotspots
Precision at 40% Recall
0.0
0.2
0.4
0.6
0.8
-t Sample 0
-t Sample 1
-t Sample 2
-t Sample 3
-t Sample 4
Mean

FIG. 24

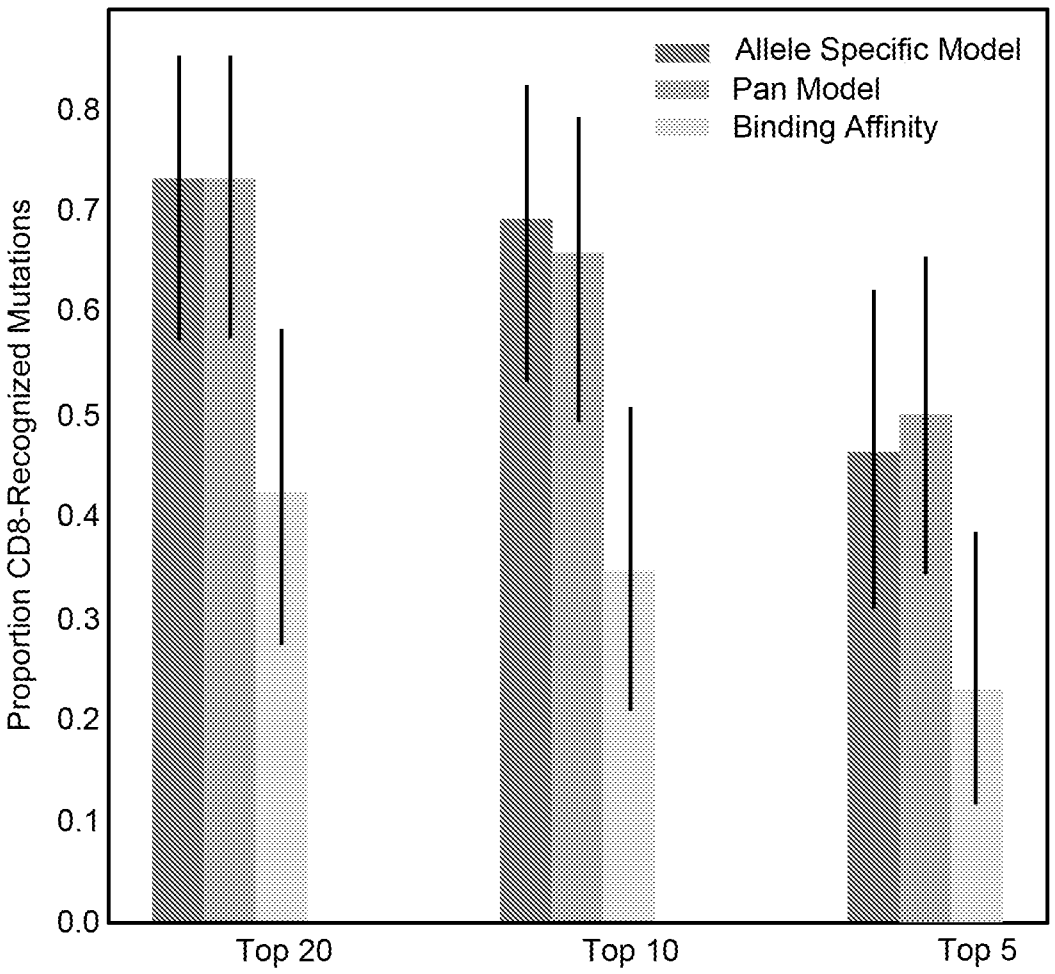

Allele Specific Model
Pan Model
Binding Affinity
Proportion CD8-Recognized Mutations
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Top 20
Top 10
Top 5

FIG. 25A

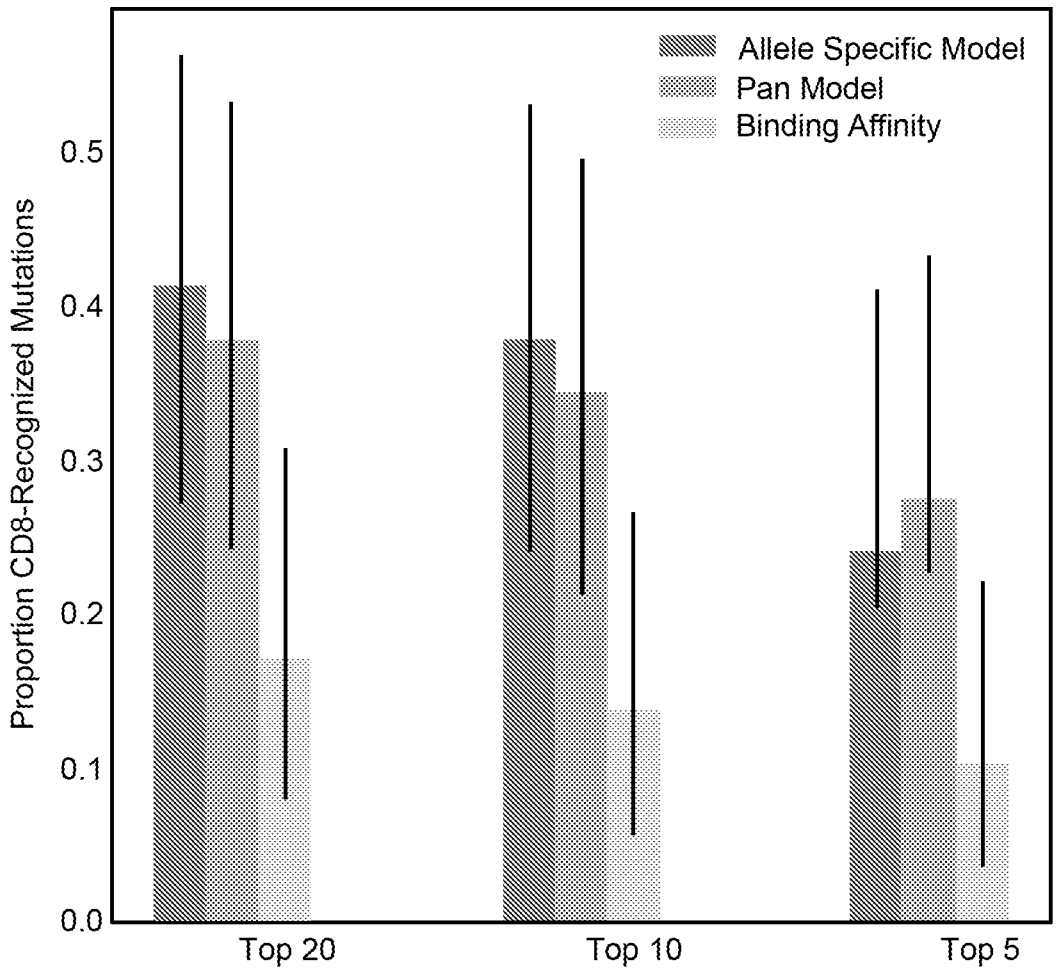

Allele Specific Model
Pan Model
Binding Affinity
Proportion CD8-Recognized Mutations
0.0
0.1
0.2
0.3
0.4
0.5
Top 20
Top 10
Top 5

FIG. 25B

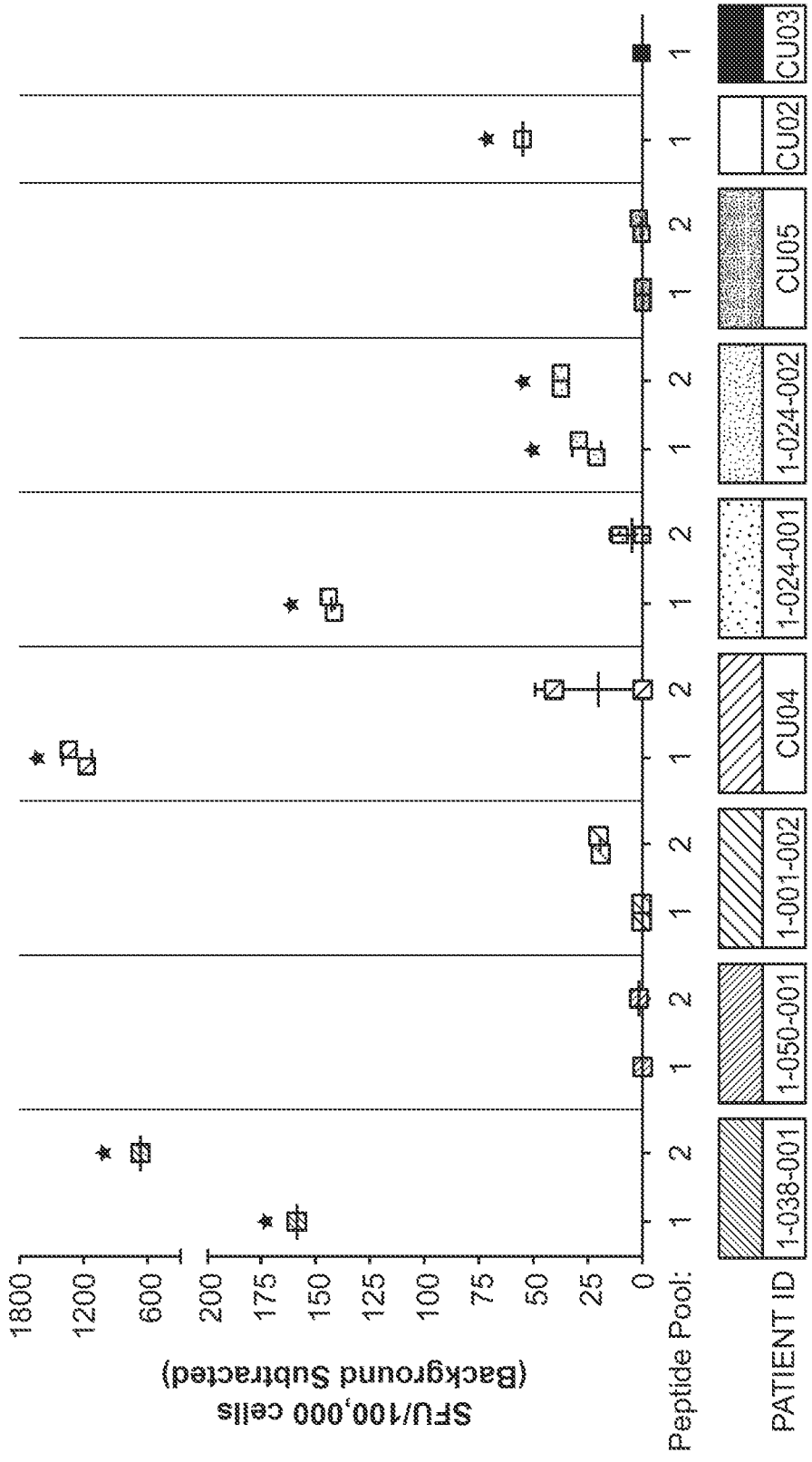

SFU/100,000 cells
(Background Subtracted)
1800
1200
600
200
175
150
125
100
75
50
25
0
Peptide Pool:
1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1
PATIENT ID
1-038-001
1-050-001
1-001-002
CU04
1-024-001
1-024-002
CU05
CU02
CU03
ELISpot stimulation (Patient-specific peptide pool #)

FIG. 26A

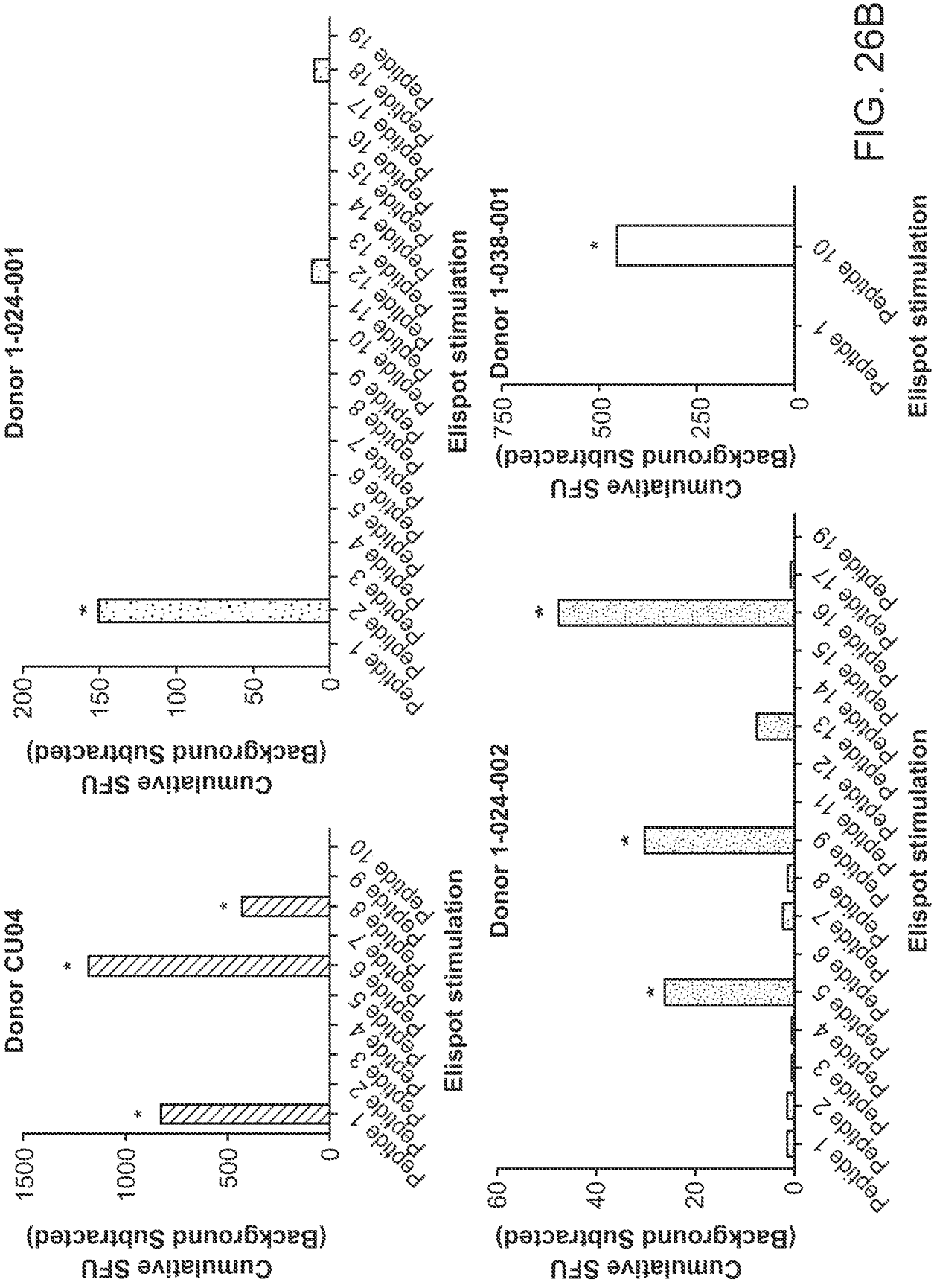

Donor CU04
Cumulative SFU (Background Subtracted)
1500
1000
500
0
Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide 10
Elispot stimulation
Donor 1-024-001
Cumulative SFU (Background Subtracted)
200
150
100
50
0
Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide 10
Peptide 11
Peptide 12
Peptide 13
Peptide 14
Peptide 15
Peptide 16
Peptide 17
Peptide 18
Peptide 19
Elispot stimulation
Donor 1-024-002
Cumulative SFU (Background Subtracted)
60
40
20
0
Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide 11
Peptide 12
Peptide 13
Peptide 14
Peptide 15
Peptide 16
Peptide 17
Peptide 19
Elispot stimulation
Donor 1-038-001
Cumulative SFU (Background Subtracted)
750
500
250
0
Peptide 1
Peptide 10
Elispot stimulation

FIG. 26B

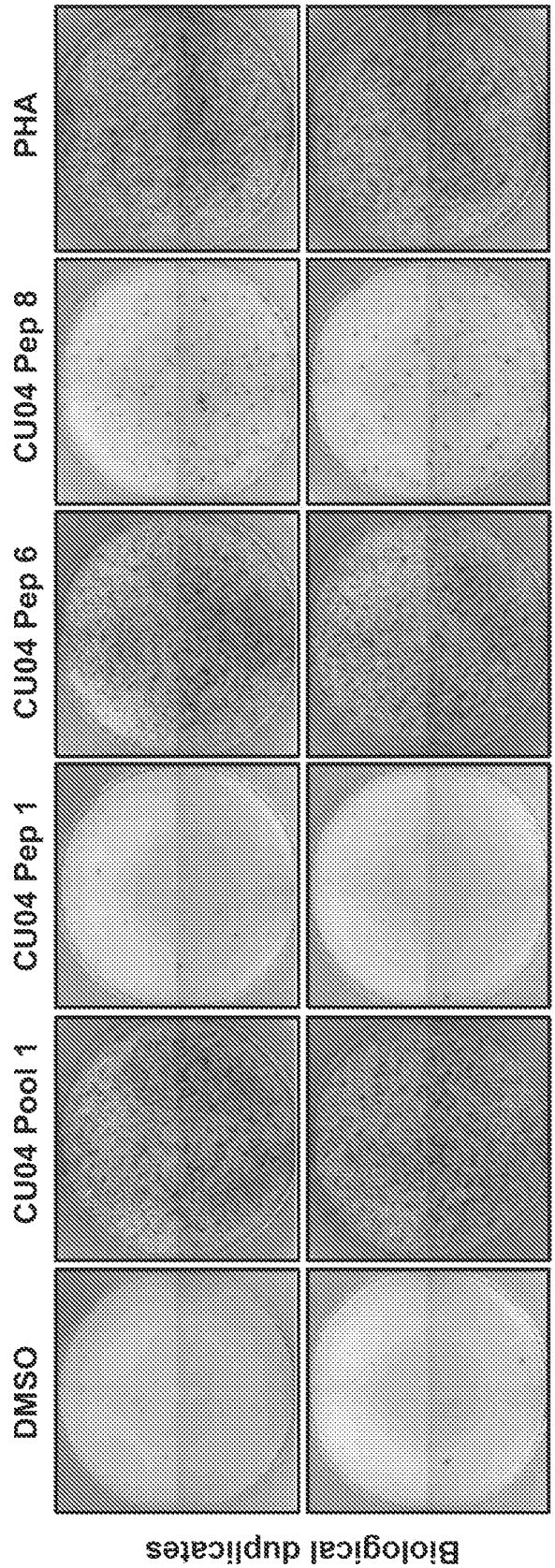

DMSO
CU04 Pool 1
CU04 Pep 1
CU04 Pep 6
CU04 Pep 8
PHA
Biological duplicates

FIG. 26C

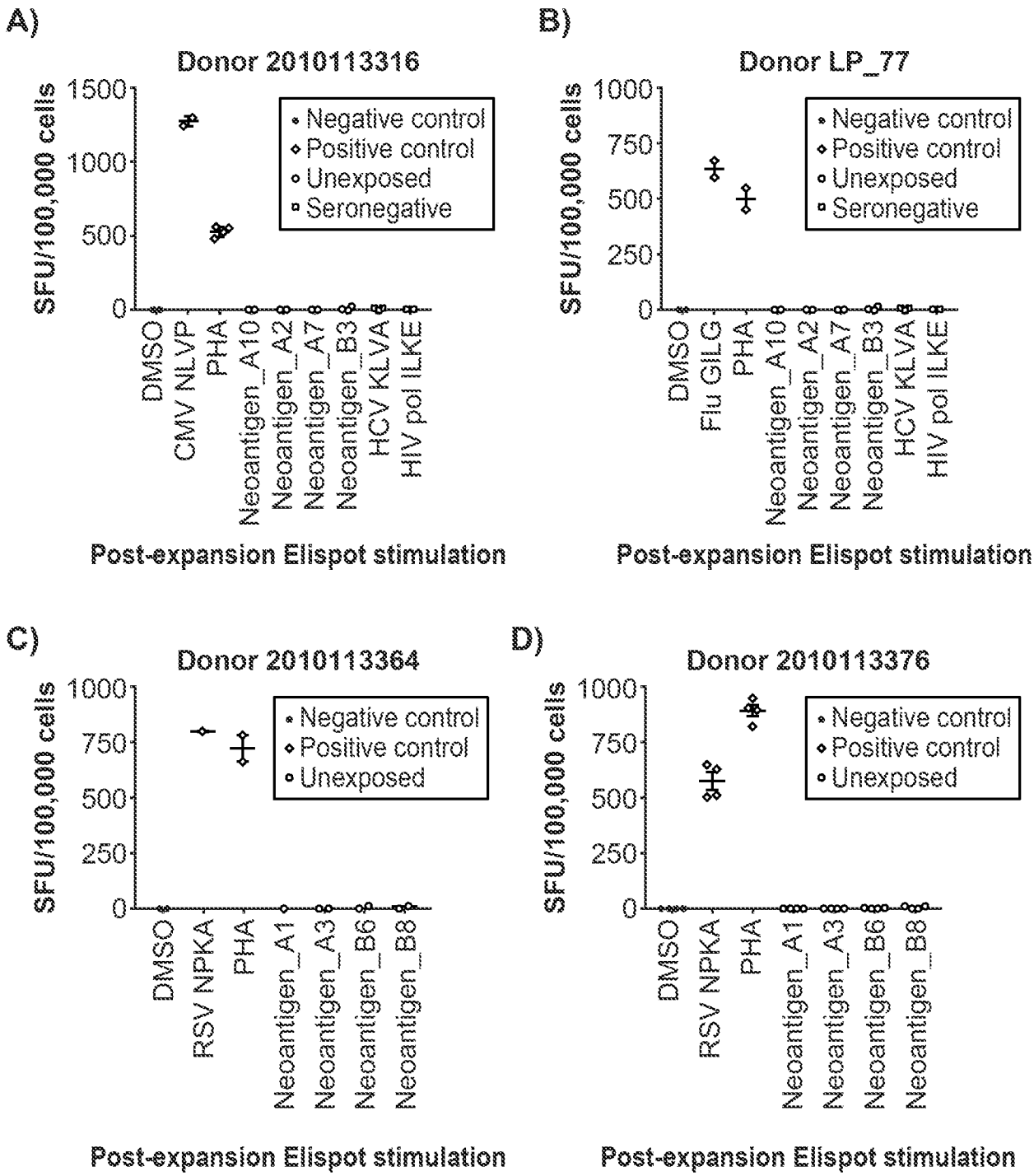

A)
Donor 2010113316
SFU/100,000 cells
1500
1000
500
0
Negative control
Positive control
Unexposed
Seronegative
DMSO
CMV NLVP
PHA
Neoantigen_A10
Neoantigen_A2
Neoantigen_A7
Neoantigen_B3
HCV KLVA
HIV pol ILKE
Post-expansion Elispot stimulation
B)
Donor LP_77
SFU/100,000 cells
1000
750
500
250
0
Negative control
Positive control
Unexposed
Seronegative
DMSO
Flu GILG
PHA
Neoantigen_A10
Neoantigen_A2
Neoantigen_A7
Neoantigen_B3
HCV KLVA
HIV pol ILKE
Post-expansion Elispot stimulation
C)
Donor 2010113364
SFU/100,000 cells
1000
750
500
250
0
Negative control
Positive control
Unexposed
DMSO
RSV NPKA
PHA
Neoantigen_A1
Neoantigen_A3
Neoantigen_B6
Neoantigen_B8
Post-expansion Elispot stimulation
D)
Donor 2010113376
SFU/100,000 cells
1000
750
500
250
0
Negative control
Positive control
Unexposed
DMSO
RSV NPKA
PHA
Neoantigen_A1
Neoantigen_A3
Neoantigen_B6
Neoantigen_B8
Post-expansion Elispot stimulation

FIG. 27A

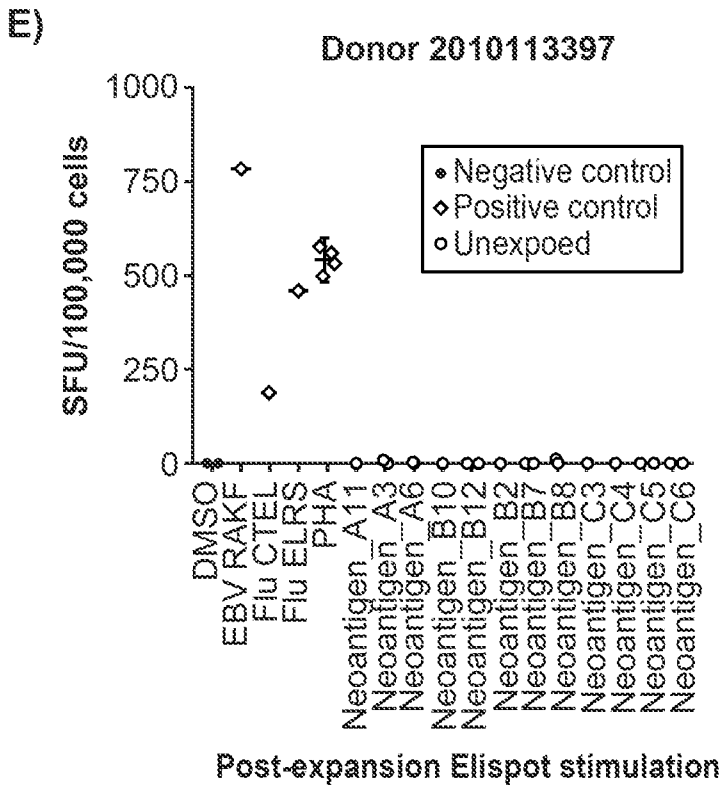

E)
Donor 2010113397
SFU/100,000 cells
1000
750
500
250
0
Negative control
Positive control
Unexpoed
DMSO
EBV RAKF
Flu CTEL
Flu ELRS
PHA
Neoantigen_A11
Neoantigen_A3
Neoantigen_A6
Neoantigen_B10
Neoantigen_B12
Neoantigen_B2
Neoantigen_B7
Neoantigen_B8
Neoantigen_C3
Neoantigen_C4
Neoantigen_C5
Neoantigen_C6
Post-expansion Elispot stimulation

FIG. 27A (Cont.)

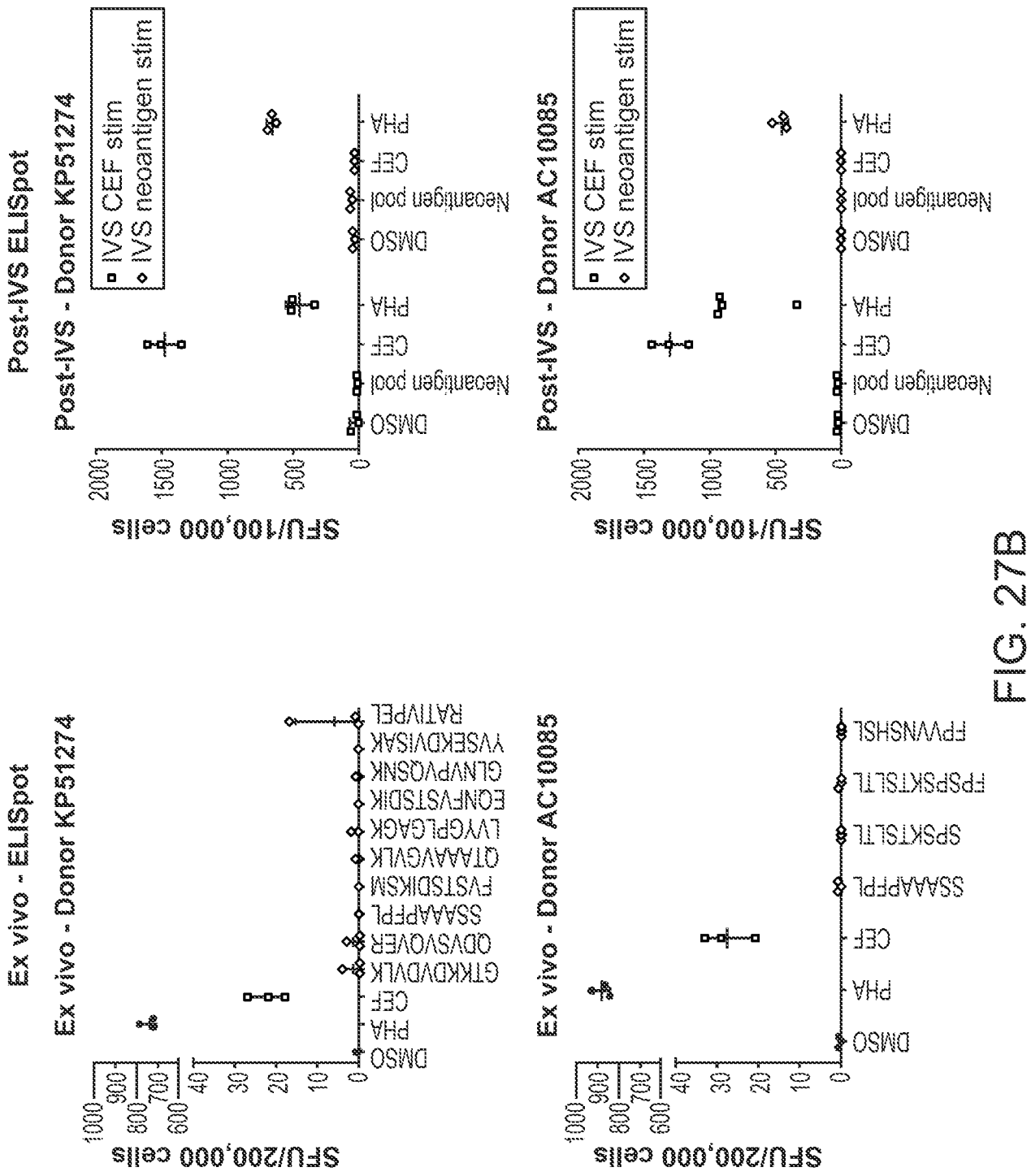

Post-IVS ELISpot
Post-IVS - Donor KP51274
IVS CEF stim
IVS neoantigen stim
SFU/100,000 cells
2000
1500
1000
500
0
DMSO
Neoantigen pool
CEF
PHA
Post-IVS - Donor AC10085
Ex vivo - ELISpot
Ex vivo - Donor KP51274
SFU/200,000 cells
1000
900
800
700
600
40
30
20
10
0
DMSO
PHA
CEF
GTKKDVDVLK
QDVSVQVER
SSAAAPFPL
FVSTSDIKSM
QTAAAVGVLK
LVYGPLGAGK
EGNFVSTSDIK
GLNVPVQSNK
YVSEKDVISAK
RATIVPEL
Ex vivo - Donor AC10085
SSAAAPFPL
SPSKTSLTL
FPSPSKTSLTL
FPVVNSHSL

FIG. 27B

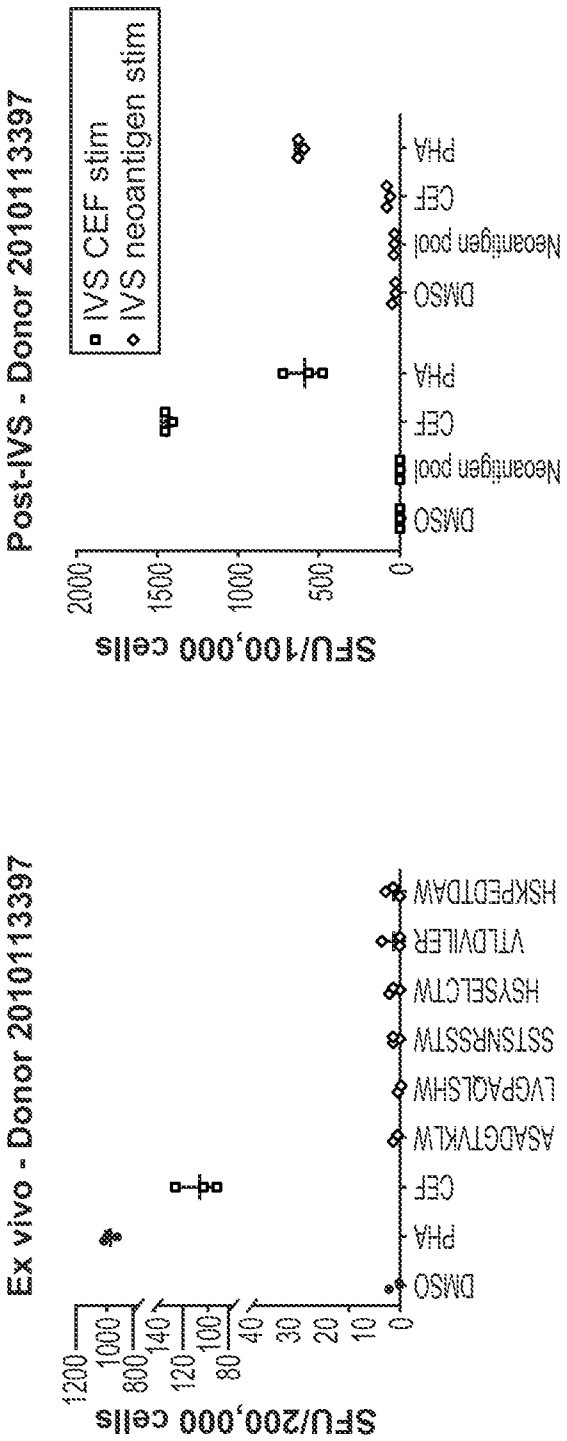

Post-IVS - Donor 20101133397
IVS CEF stim
IVS neoantigen stim
SFU/100,000 cells
2000
1500
1000
500
0
DMSO
Neoantigen pool
CEF
PHA
DMSO
Neoantigen pool
CEF
PHA
Ex vivo - Donor 20101133397
SFU/200,000 cells
1200
1000
800
140
120
100
80
40
30
20
10
0
DMSO
PHA
CEF
ASADGTVKLW
LVGPAQLSHW
SSTSNRSSTW
HSYSELCTW
VTLDVILER
HSKPEDTDAW

FIG. 27B (Cont.)

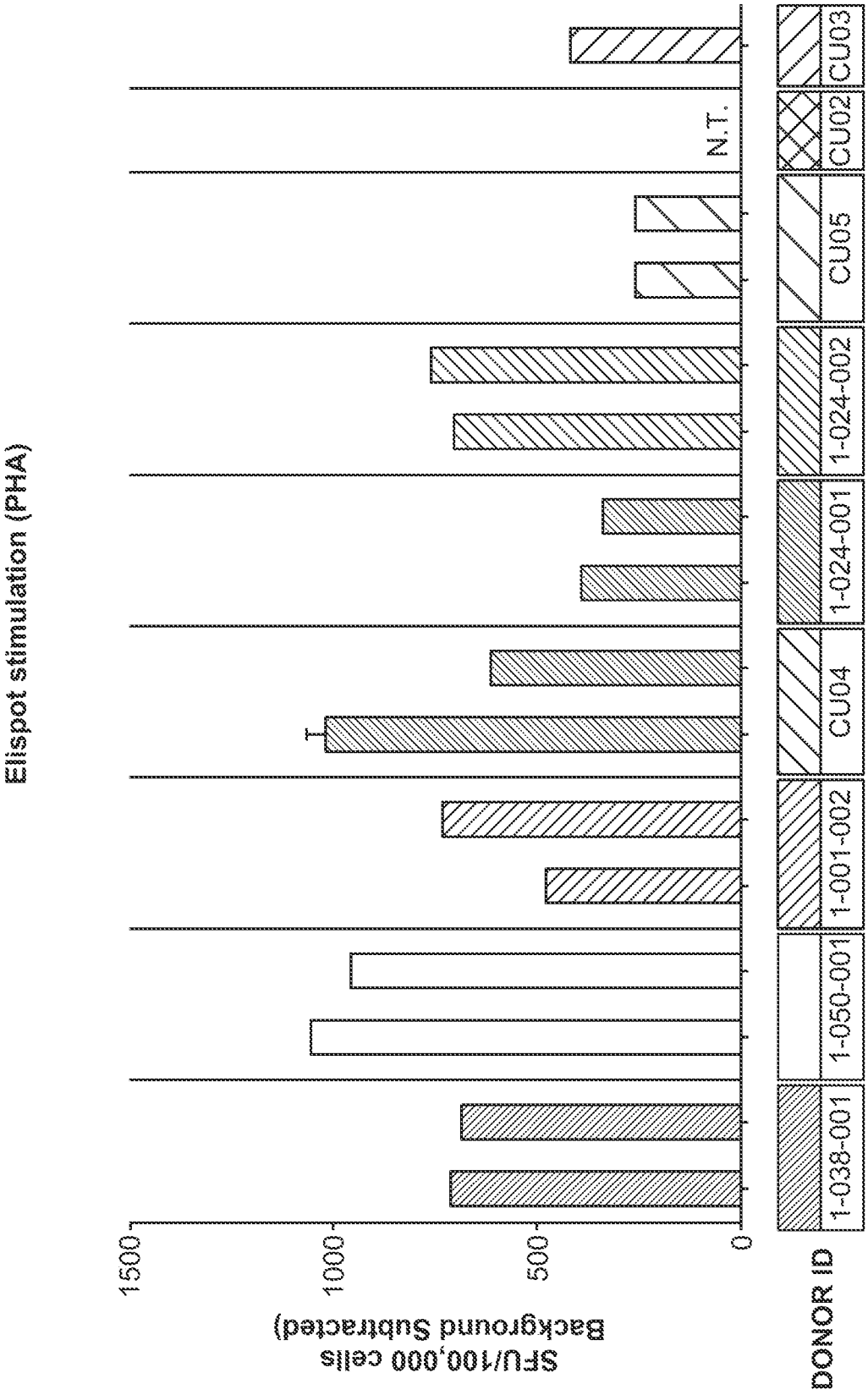

Elispot stimulation (PHA)
SFU/100,000 cells
Background Subtracted)
1500
1000
500
0
T
N.T.
DONOR ID
1-038-001
1-050-001
1-001-002
CU04
1-024-001
1-024-002
CU05
CU02
CU03

FIG. 28

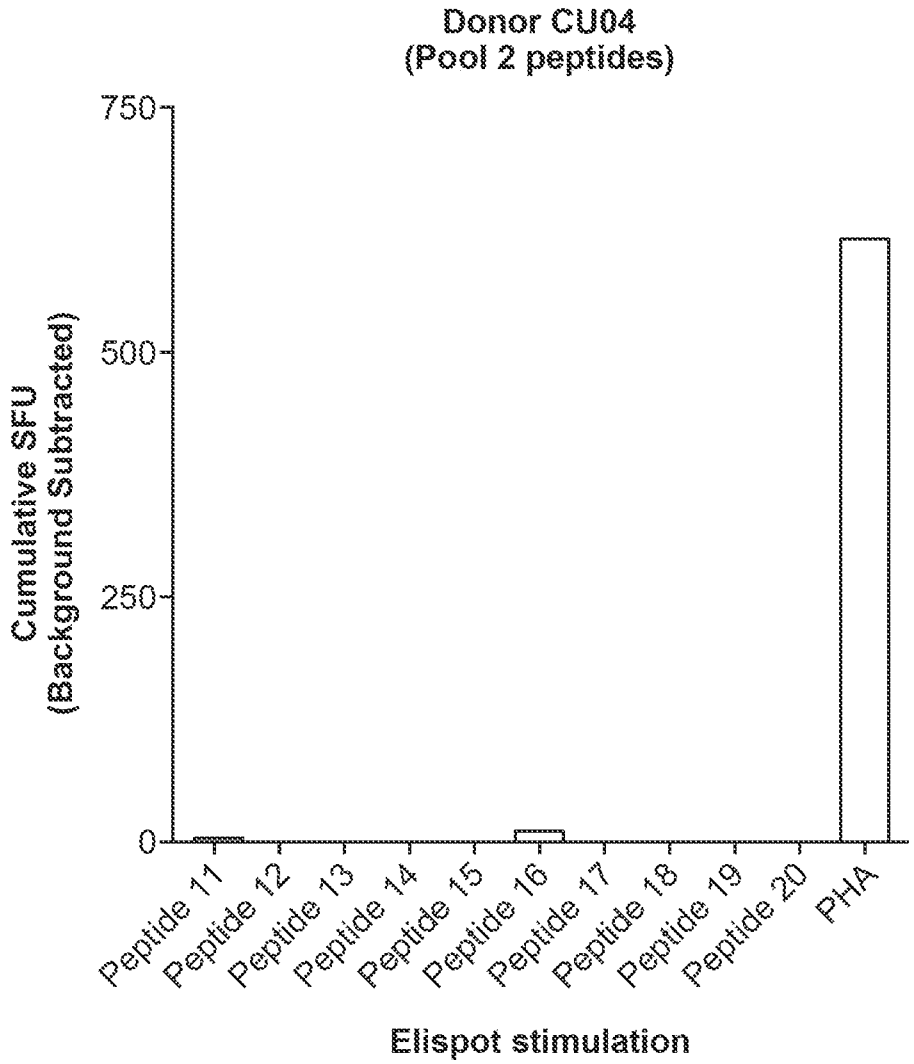

Donor CU04
(Pool 2 peptides)
Cumulative SFU
(Background Subtracted)
750
500
250
0
Peptide 11
Peptide 12
Peptide 13
Peptide 14
Peptide 15
Peptide 16
Peptide 17
Peptide 18
Peptide 19
Peptide 20
PHA
Elispot stimulation

FIG. 29A

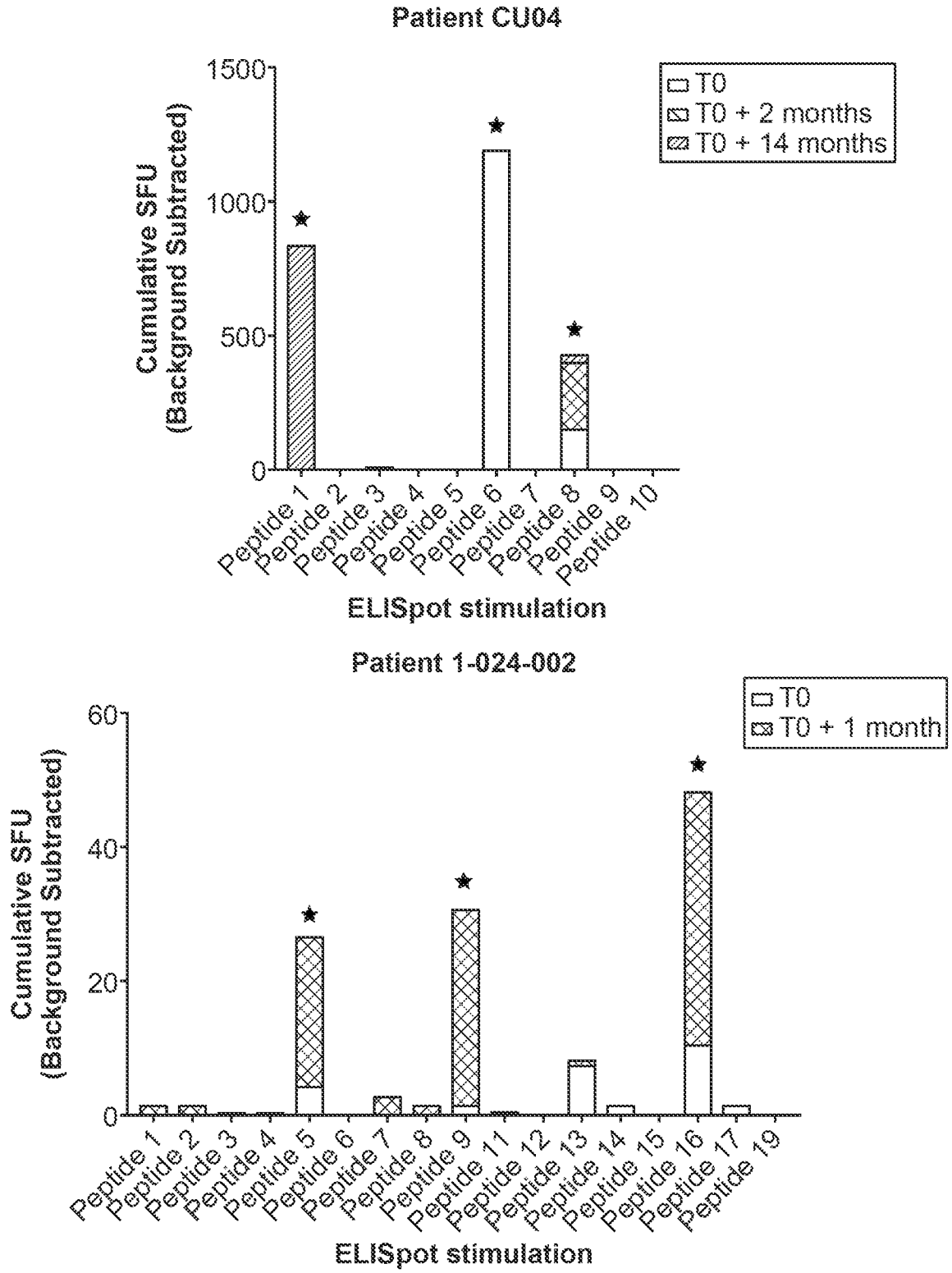

Patient CU04
Cumulative SFU
(Background Subtracted)
1500
1000
500
0
T0
T0 + 2 months
T0 + 14 months
Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide 10
ELISpot stimulation
Patient 1-024-002
Cumulative SFU
(Background Subtracted)
60
40
20
0
T0
T0 + 1 month
Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide 11
Peptide 12
Peptide 13
Peptide 14
Peptide 15
Peptide 16
Peptide 17
Peptide 19
ELISpot stimulation

FIG. 29B

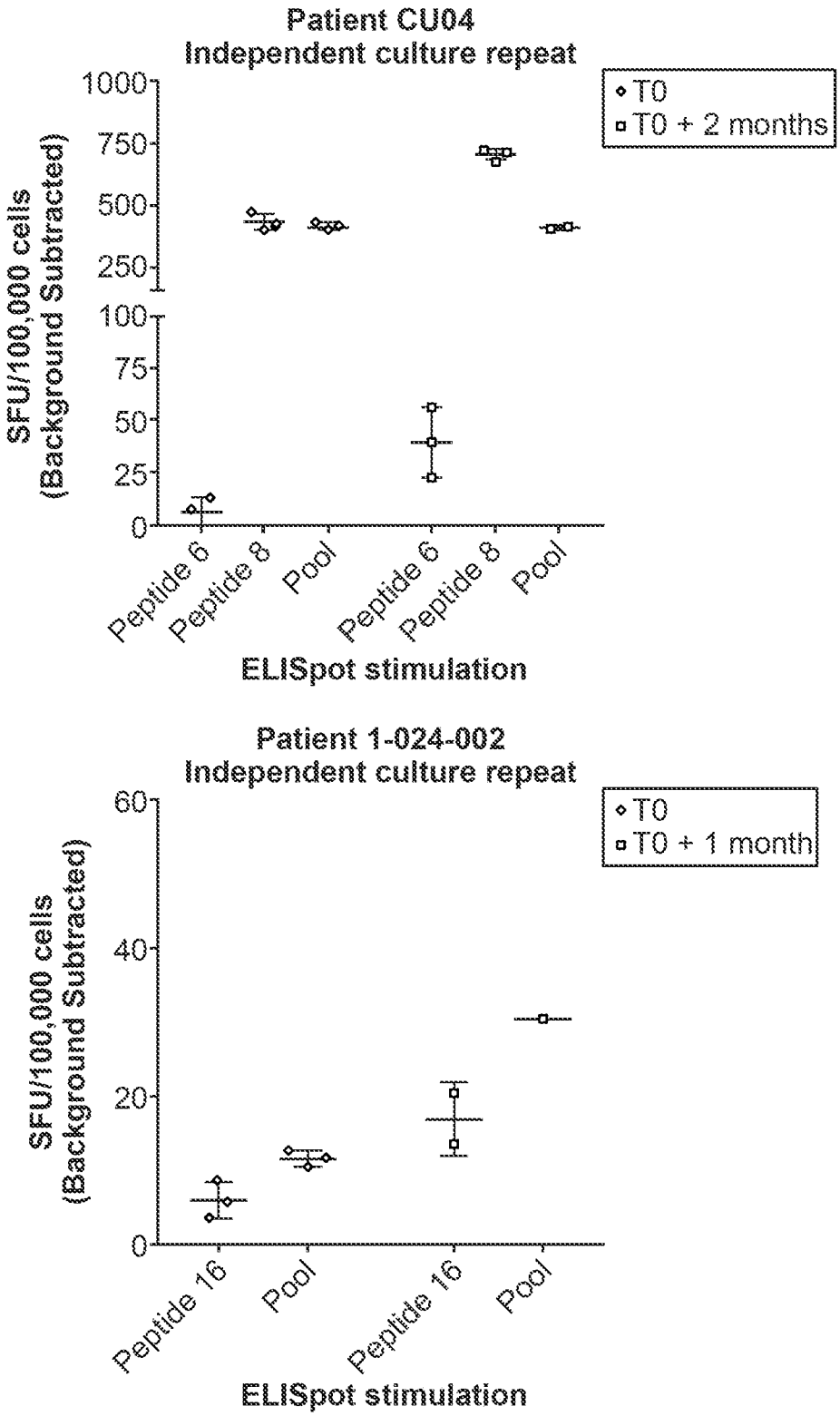

Patient CU04
Independent culture repeat
T0
T0 + 2 months
SFU/100,000 cells
(Background Subtracted)
1000
750
500
250
100
75
50
25
0
Peptide 6
Peptide 8
Pool
Peptide 6
Peptide 8
Pool
ELISpot stimulation
Patient 1-024-002
Independent culture repeat
T0
T0 + 1 month
SFU/100,000 cells
(Background Subtracted)
60
40
20
0
Peptide 16
Pool
Peptide 16
Pool
ELISpot stimulation

FIG. 29C

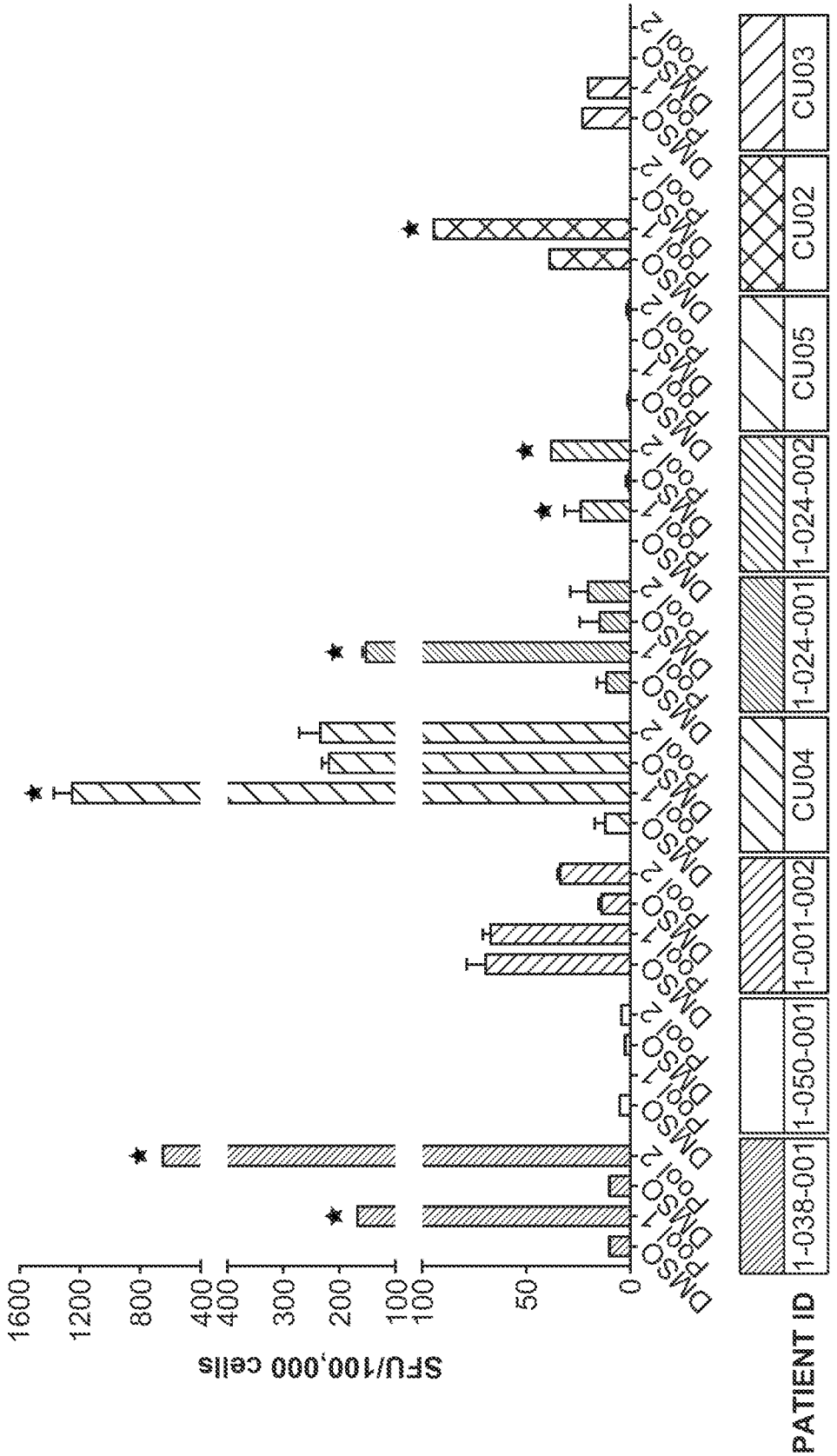

SFU/100,000 cells
1600
1200
800
400
400
300
200
100
100
50
0
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
DMSO 1 Pool 2
DMSO 2 Pool 2
PATIENT ID
1-038-001
1-050-001
1-001-002
CU04
1-024-001
1-024-002
CU05
CU02
CU03

FIG. 30

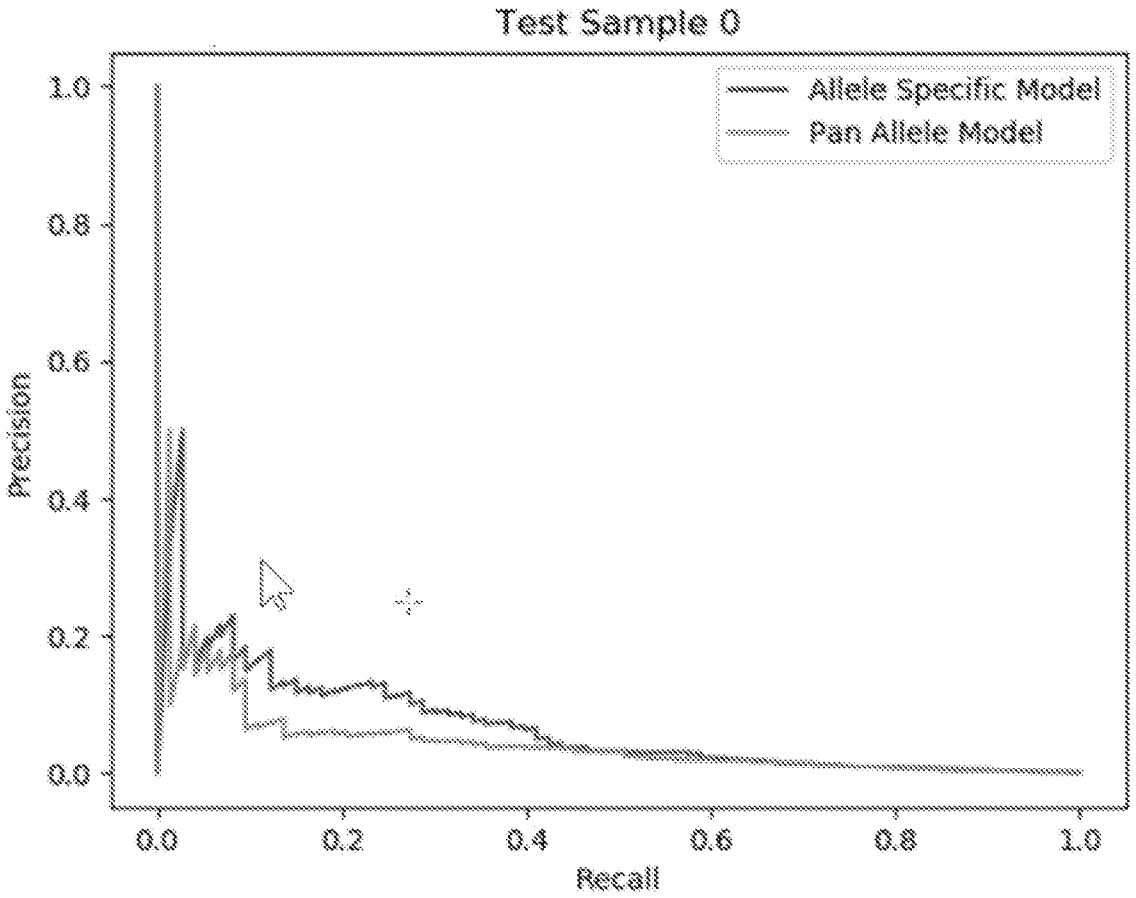

Test Sample 0
Allele Specific Model
Pan Allele Model
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 31A

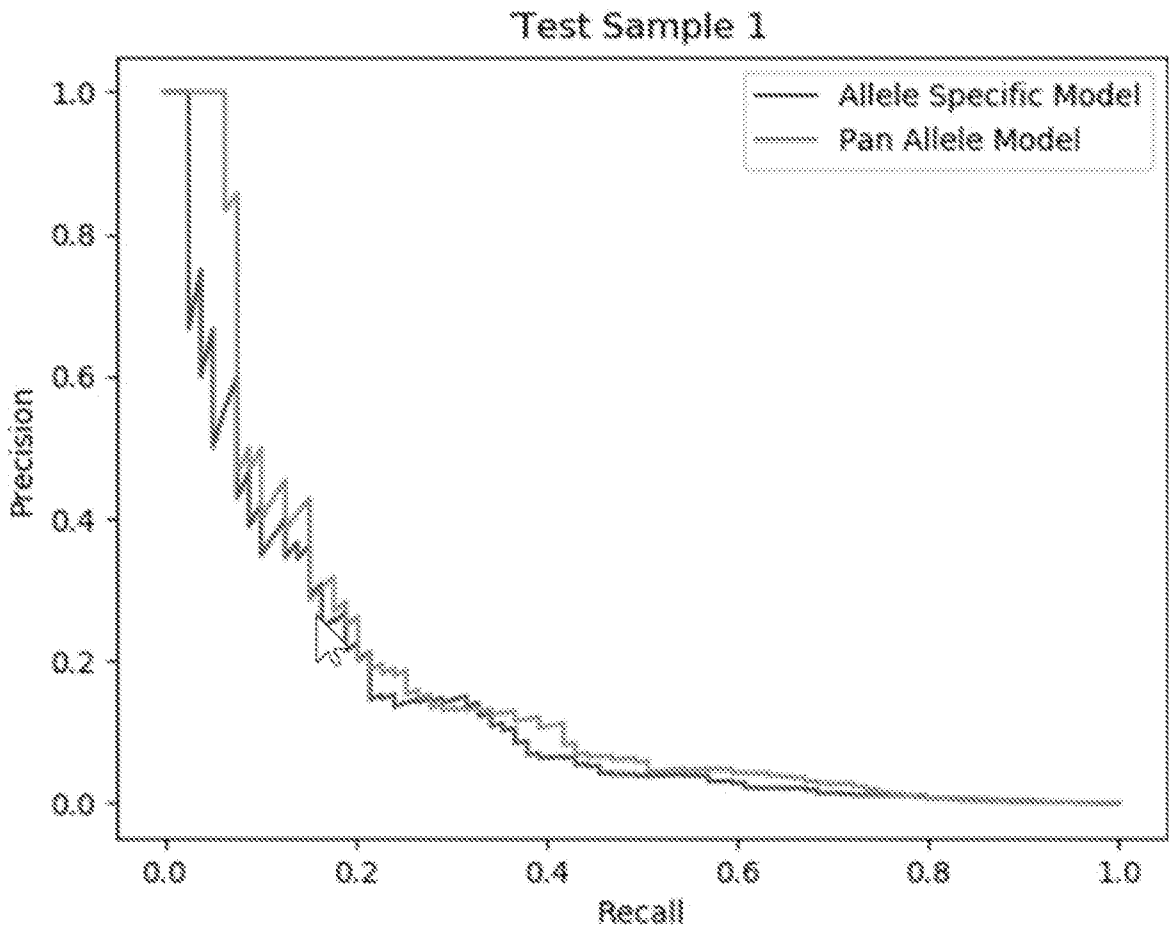

Test Sample 1
Allele Specific Model
Pan Allele Model
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 31B

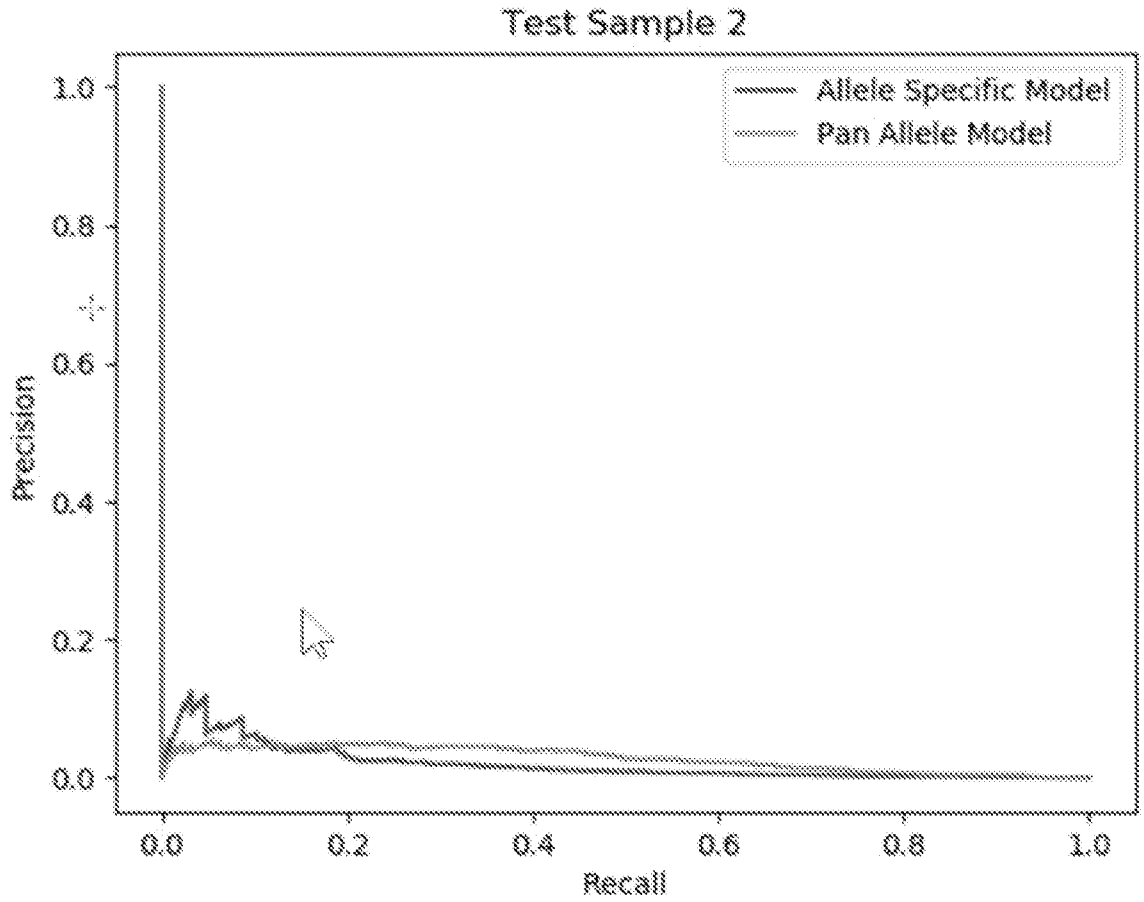

Test Sample 2
Allele Specific Model
Pan Allele Model
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

FIG. 31C

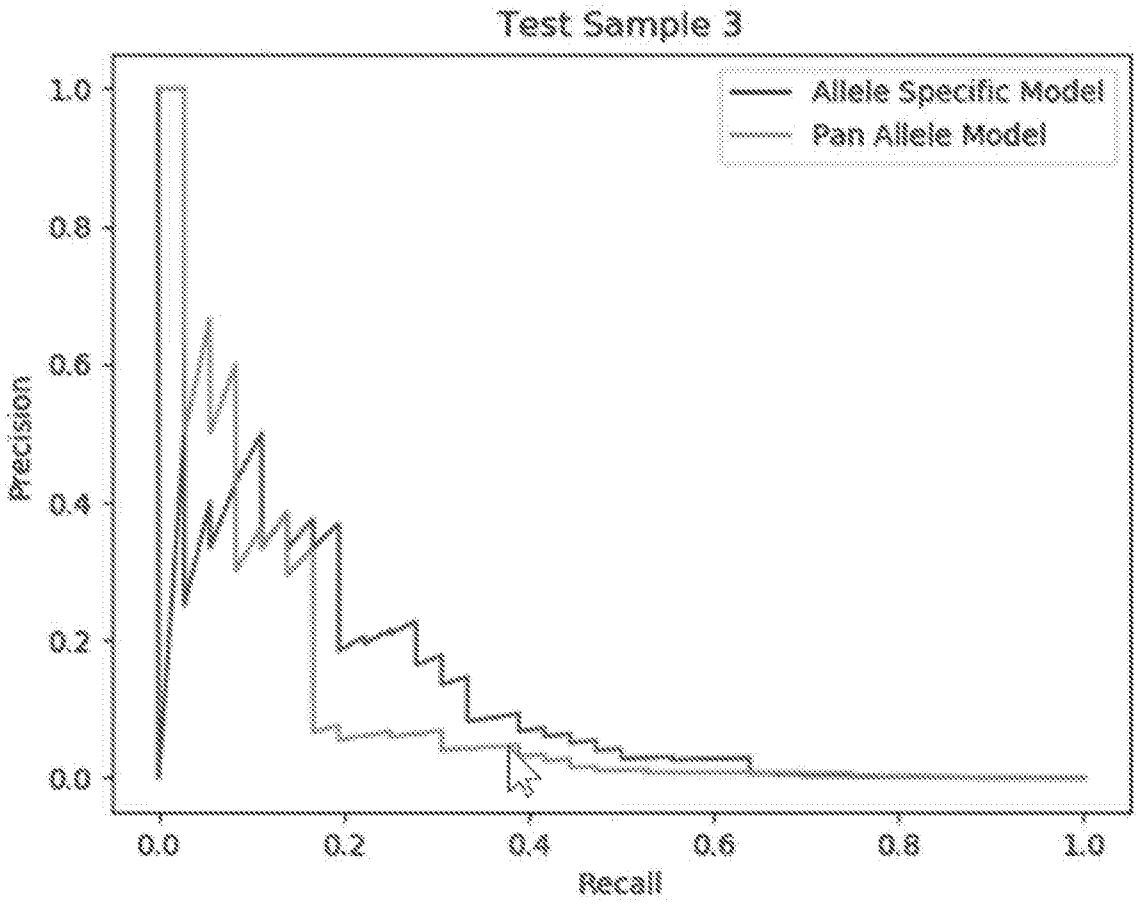

Test Sample 3
Allele Specific Model
Pan Allele Model
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 31D

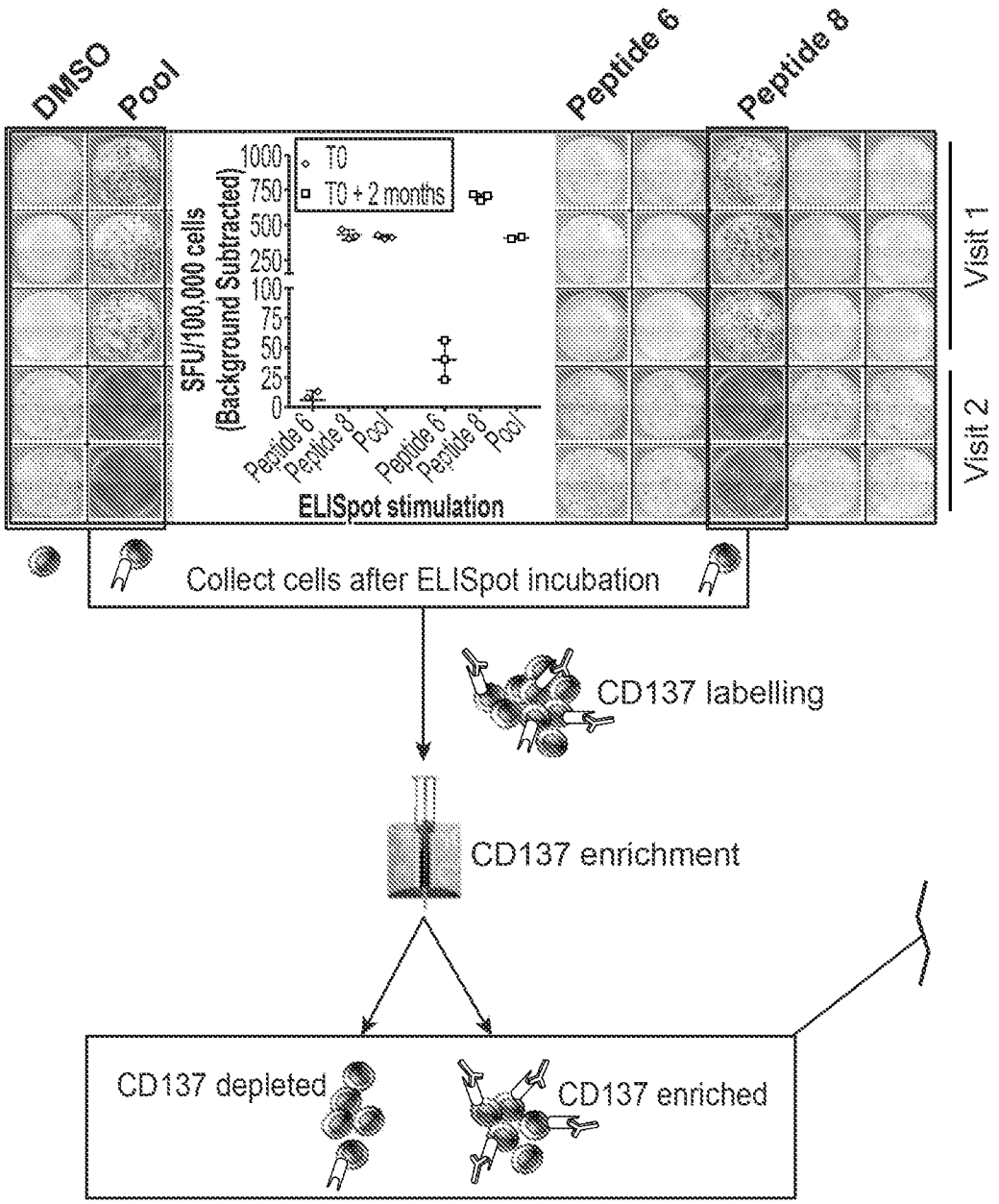

DMSO
Pool
Peptide 6
Peptide 8
Visit 1
Visit 2
T0
T0 + 2 months
SFU/100,000 cells
(Background Subtracted)
1000
750
500
250
100
75
50
25
0
Peptide 6
Peptide 8
Pool
Peptide 6
Peptide 8
Pool
ELISpot stimulation
Collect cells after ELISpot incubation
CD137 labelling
CD137 enrichment
CD137 depleted
CD137 enriched

FIG. 32

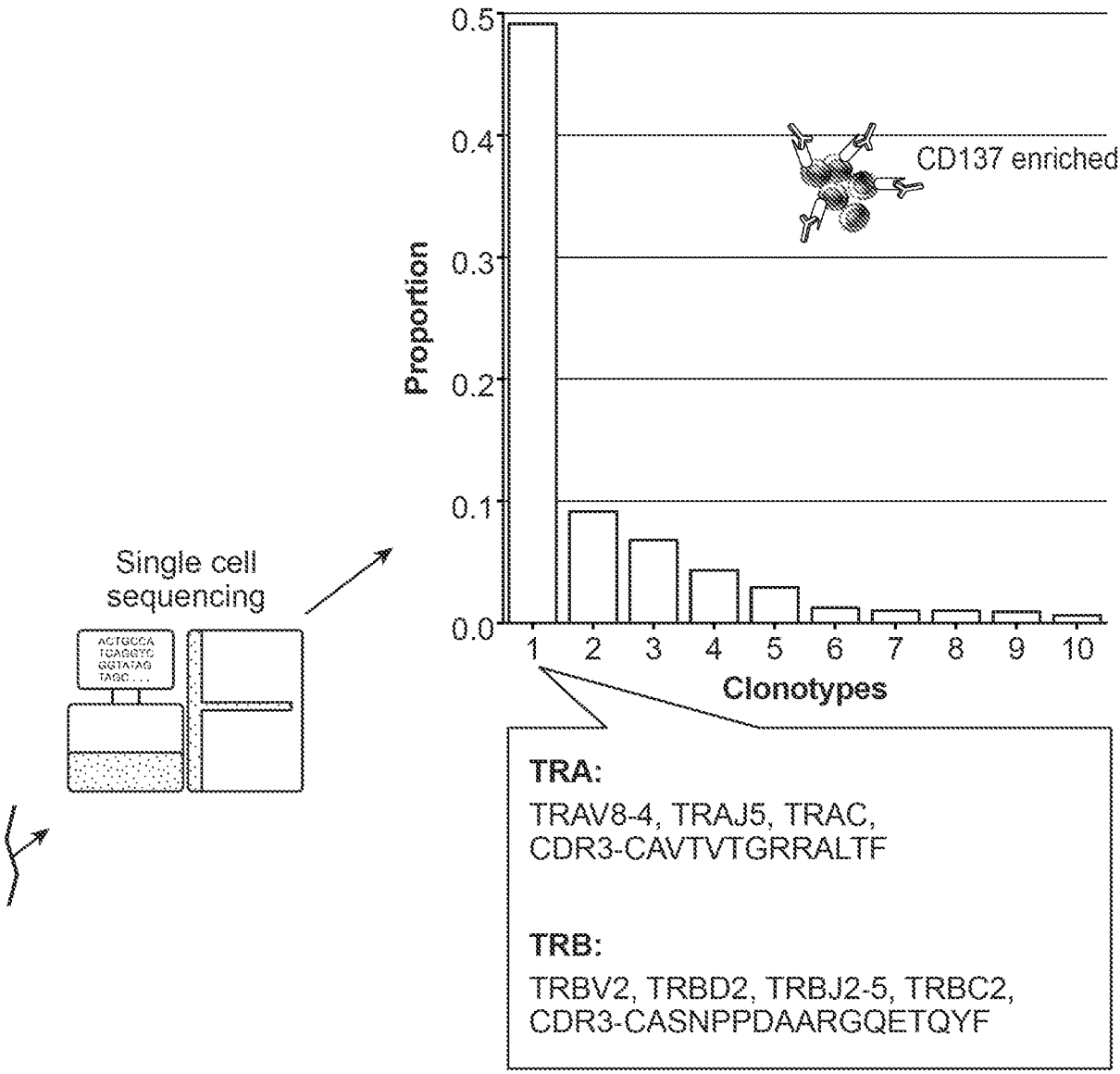

Single cell sequencing
ACTGCCA
TGAGGTC
GGTATAG
TAGC . . .
Proportion
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10
Clonotypes
CD137 enriched
TRA:
TRAV8-4, TRAJ5, TRAC,
CDR3-CAVTVTGRRALTF
TRB:
TRBV2, TRBD2, TRBJ2-5, TRBC2,
CDR3-CASNPPDAARGQETQYF

FIG. 32 (Cont.)

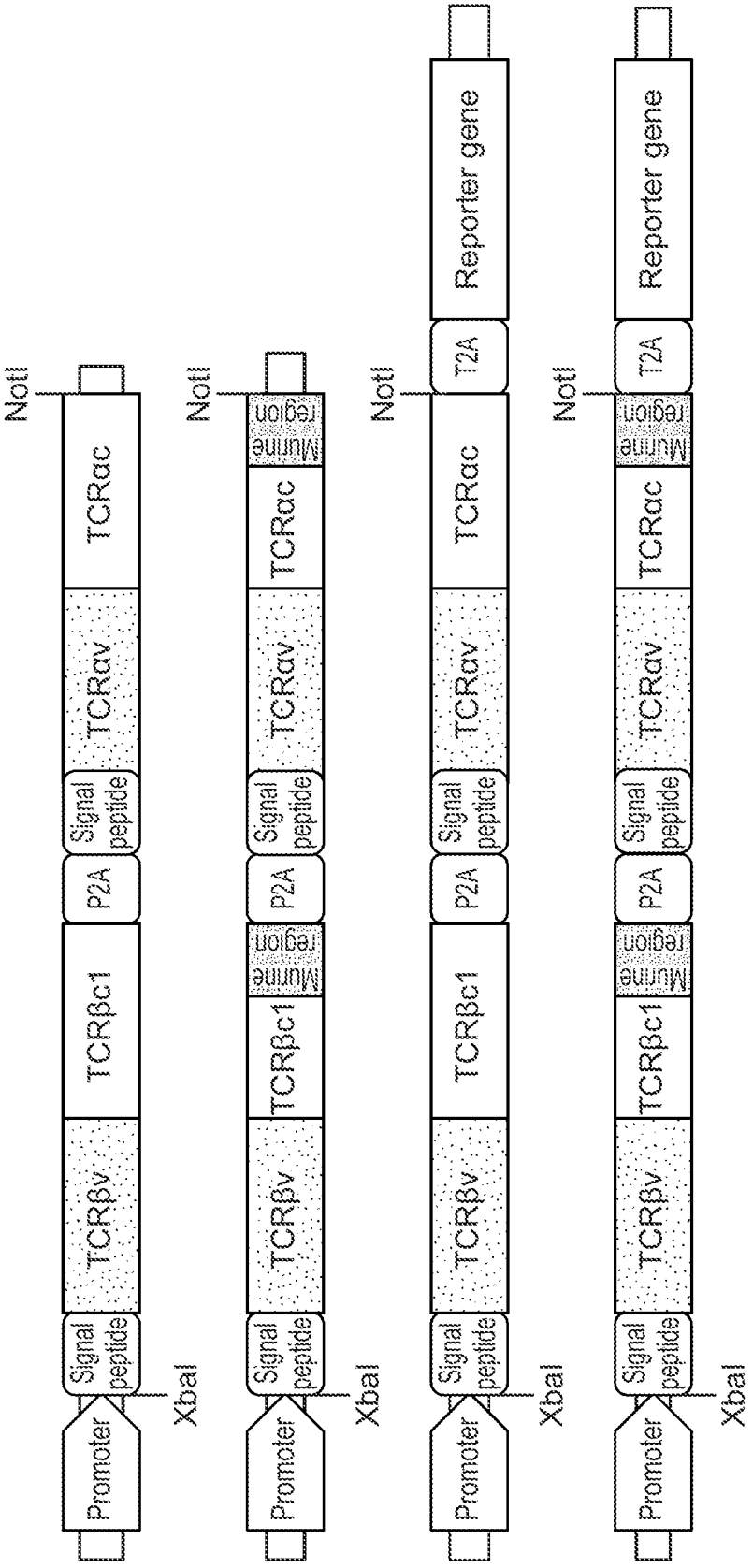

NotI
Promoter
Signal peptide
TCRβv
TCRβc1
Murine region
P2A
Signal peptide
TCRαv
TCRαc
T2A
Reporter gene
XbaI

FIG. 33

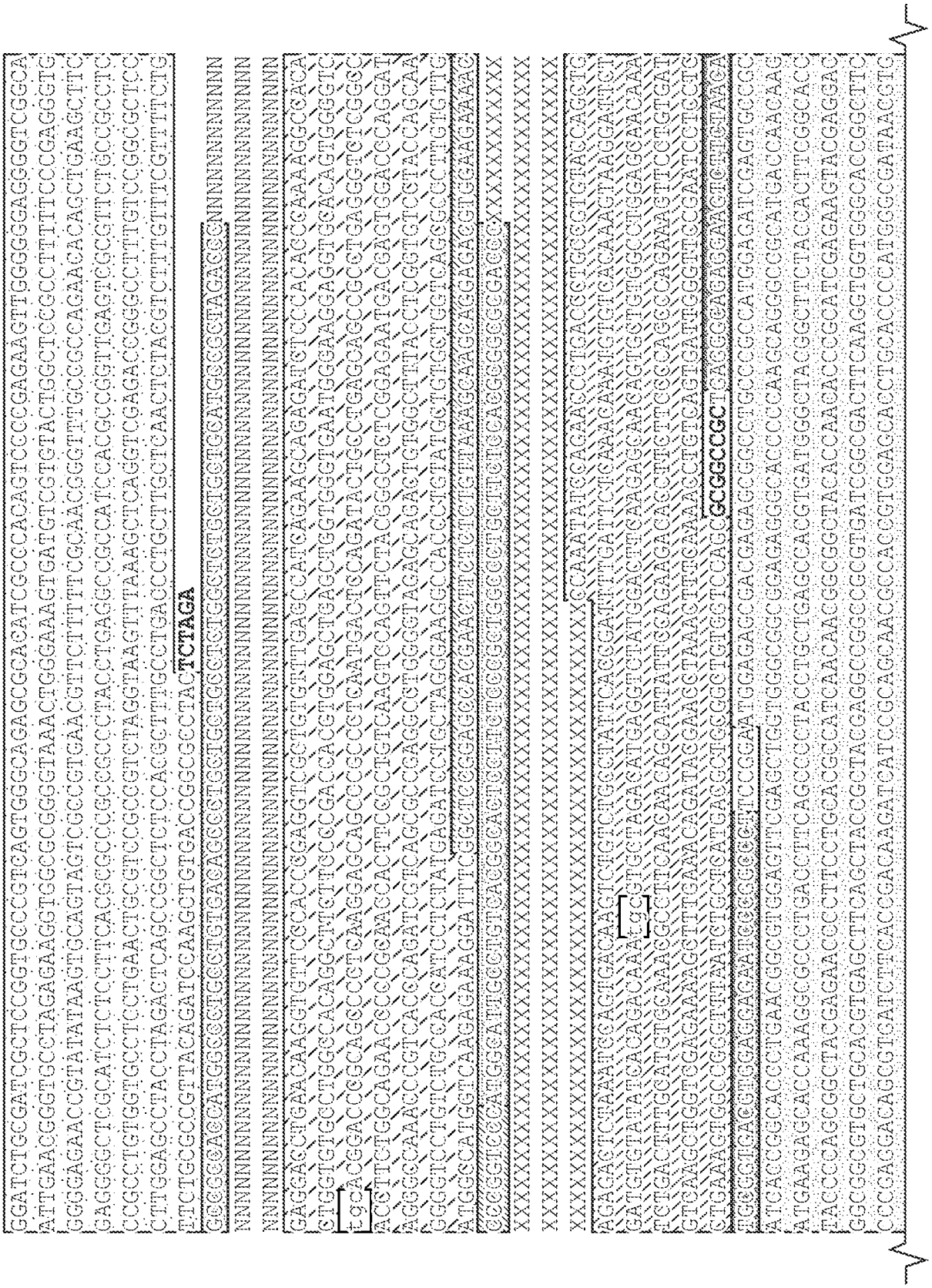
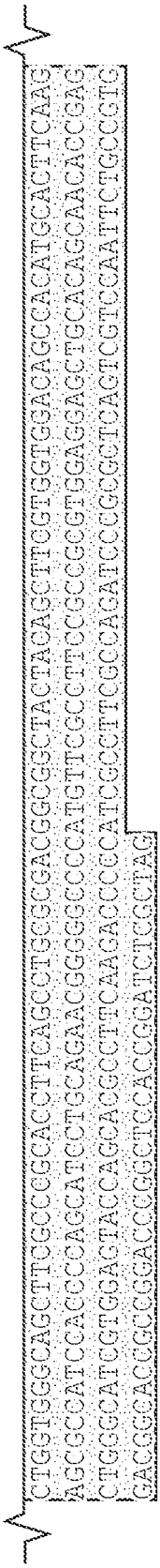


FIG. 34


CTGGTGGGCAGCCTTCGGCCCGGCACCTTCAGCCTGGCGGACGGCGGCTACTACAGCTTCGTGGTGGACAGCCACATGCACTTCAAG
AGCGCCATCCACCCCAGCATCCTGCAGAACGGGGGCCCCATGTTCGCCTTCCGCCGCGTGGAGGAGCTGCACAGCAACACCGAG
CTGGGCATCGTGGAGTACCAGCACGCCTTCAAGACCCCCATCGCCTTCGGCGCTCAGTCGTCCAATTCTGCCCGTC
GACGGCACCGGCGACCCGGGCCTCCACCCGGATCTCGCTAG

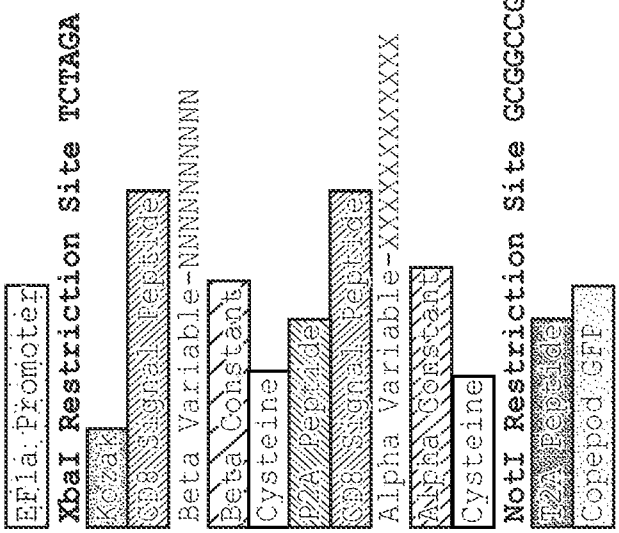

EF1a Promoter
XbaI Restriction Site TCTAGA
Kozak
Cd8 Signal Peptide
Beta Variable-NNNNNNNNN
Beta Constant
Cysteine
P2A Peptide
Cd8 Signal Peptide
Alpha Variable-XXXXXXXXXXX
Alpha Constant
Cysteine
NotI Restriction Site GCGGCCGC
P2A Peptide
Copepod GFP

FIG. 34 (Cont.)

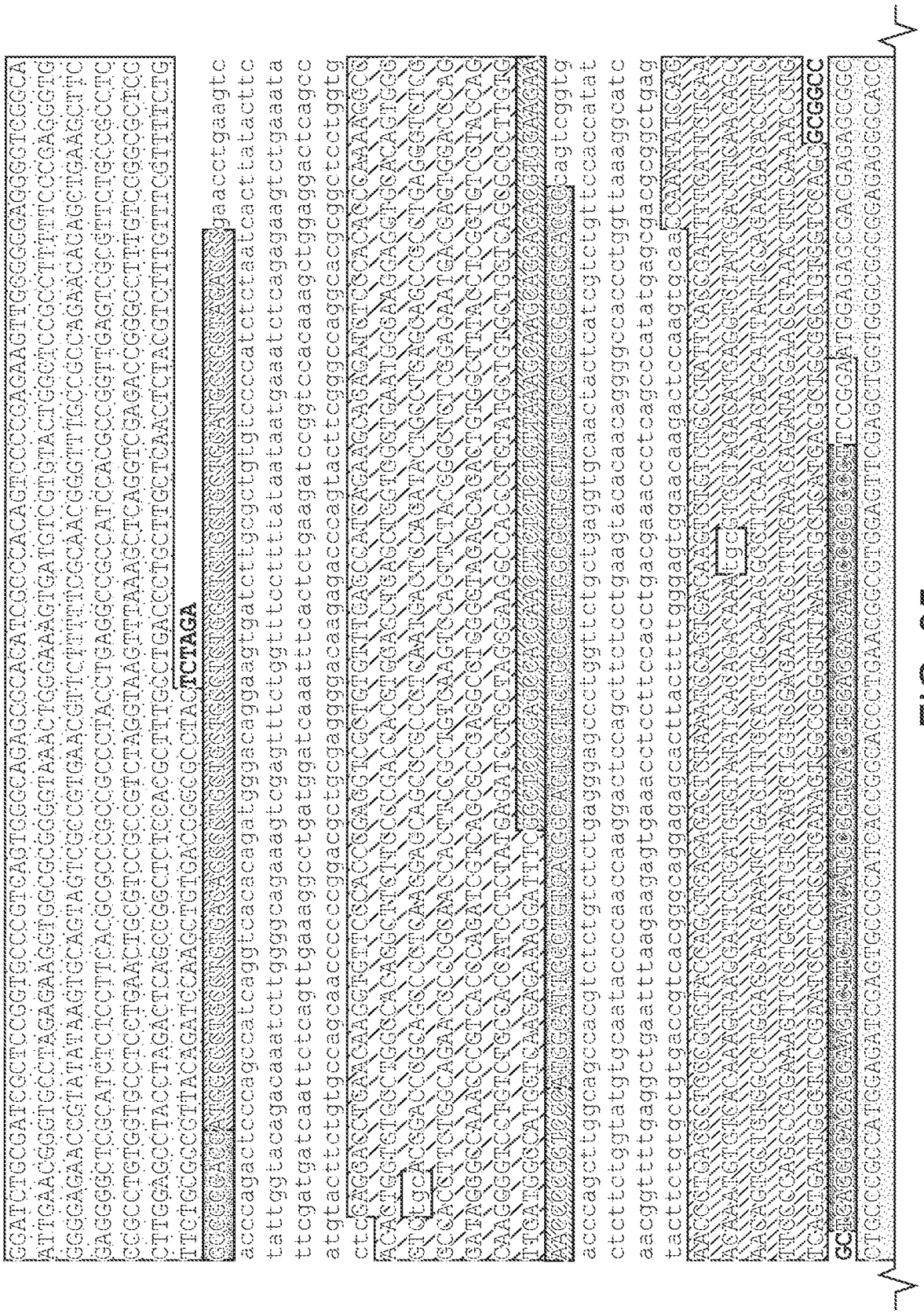
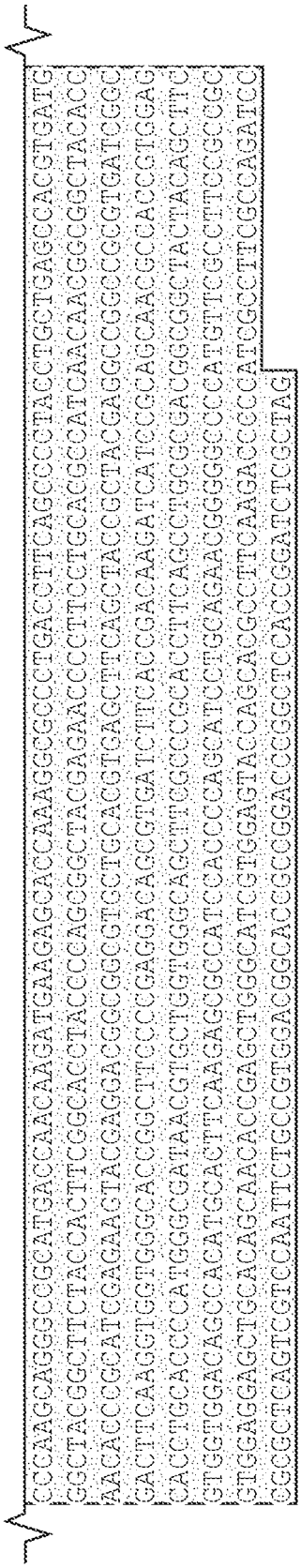


FIG. 35


CCCAAGCAGGCGCCGCATGACCAACAAGATGAAGAGCACCAAGGCGCCCTGACCTTCAGCCCCTACCTGCTGAGCCACGTGATG
GGCTACGGCTTCTACCACTTCGGCACCTACCCCAGCGGCTACGAGAACCCCTTCCTGCACGCCATCAACAACGGCGGCTACACC
AACACCCGCATCGAGAAGTACGAGGACGGCGGCGTGCTGCACGTGAGCTTCAGCTACCGCTACGAGGCCGGCCGCGTGATCGGC
GACTTCAAGGTGATGGGCACCGGCTTCCCCGAGGACAGCGTGATCTTCACCGACAAGATCATCCGCAGCAACGCCACCGTGGAG
CACCTGCACCCCATGGGCGATAACGTGCTGGTGGGCAGCTTCGCCCGCACCTTCAGCCTGCGCGACGGCGGCTACTACAGCTTC
GTGGTGGACAGCCACATGCACTTCAAGAGCGCCATCCACCCCAGCATCCTGCAGAACGGGGGCCCCATGTTCGCCTTCCGCGCC
GTGGAGGAGCTGCACAGCAACACCGAGCTGGGCATCGTGGAGTACCAGCACGCCTTCAAGACCCCCATCGCCTTCGCCAGATCC
CGCGCCTCAGTCGTCCAATTCTGCCGTGGACGGCACCGCCCGGACCCGGCTCCACCGGATCTCGCTAG

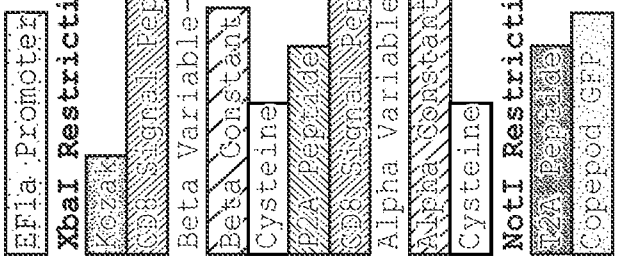

EF1a Promoter
XbaI Restriction Site TCTAGA
Kozak
CD8 Signal Peptide
Beta Variable-NNNNNNNNN
Beta Constant
Cysteine
P2A Peptide
CD8 Signal Peptide
Alpha Variable-XXXXXXXXX
Alpha Constant
Cysteine
NotI Restriction Site GCGGCCGC
T2A Peptide
Copepod GFP

FIG. 35 (Cont.)

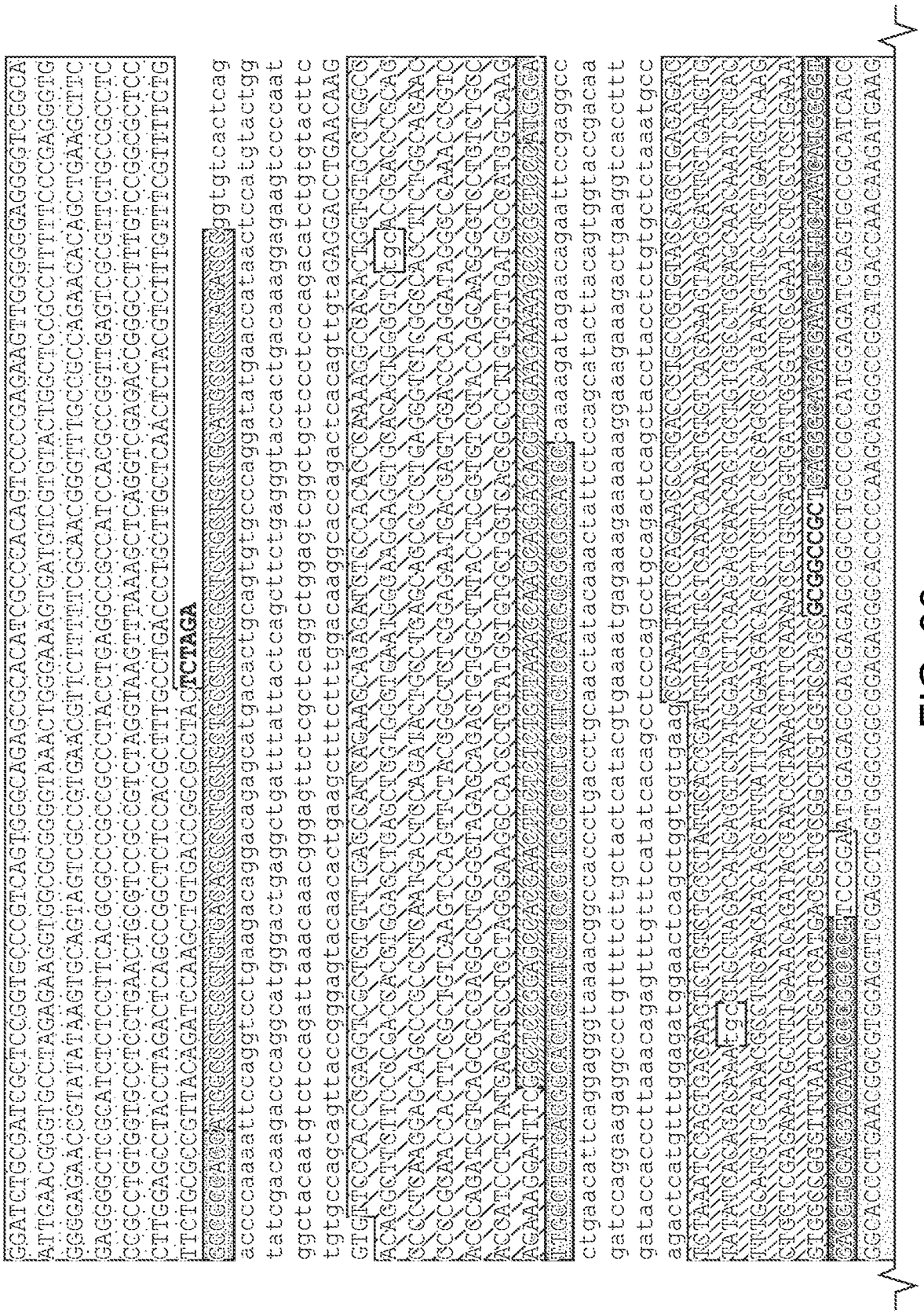
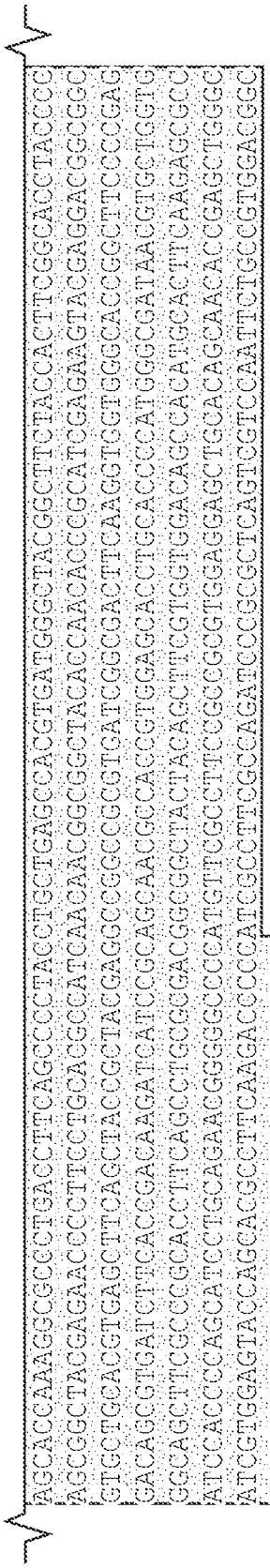


FIG. 36


AGCACCAAAGGCGCCCTGACCTTCAGCCCCTACCTGCTGAGCCACGTGATGGGCTACGGCTTCTACCACTTCGGCACCTACCCC
AGCGGCTACGAGAACCCCTTCCTGCACGCCATCAACAACCACCCCGCATCGAGAAGTACGAGGACGGCGGC
GTGCTGCACGTGAGCTTCAGCTACCGCTACGAGGGCGGCGTGATCGGCGACTTCAAGGTGGTGGGCACCGGCTTCCCCGAG
GACAGCGTGATCTTCACCGACAAGATCATCCGCAGCAACGCCACCGTGGAGCACCTGCACCCCATGGGCGATAACGTGCTGGTG
CGCAGCTTCGCCGCACCTTCAGCCTGCGGCGACGGCGGCTACTACAGCTTCGTGGTGGACAGCCACATGCACTTCAAGAGCGCC
ATCCACCCCAGCATCCTGCAGAACGGGGGCCCCATGTTCGCCTTCCGCCGCGTGGAGGAGCTGCACAGCAACACCGAGCTGGGC
ATCGTGGAGTACCAGCACGCCTTCAAGACCCCCATCGCCTTCAGTCGTCCAATTCTGCCGTGGACGGC
ACCGGCGGACCGGGCTCCACCGGATCTCGCTAG

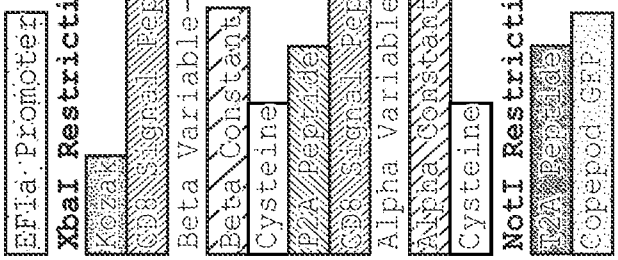

EF1a Promoter
XbaI Restriction Site TCTAGA
Kozak
CD8 Signal Peptide
Beta Variable-NNNNNNNNNN
Beta Constant
Cysteine
P2A Peptide
CD8 Signal Peptide
Alpha Variable-XXXXXXXXXX
Alpha Constant
Cysteine
NotI Restriction Site GCGGCCGC
T2A Peptide
Copepod GFP

FIG. 36 (Cont.)

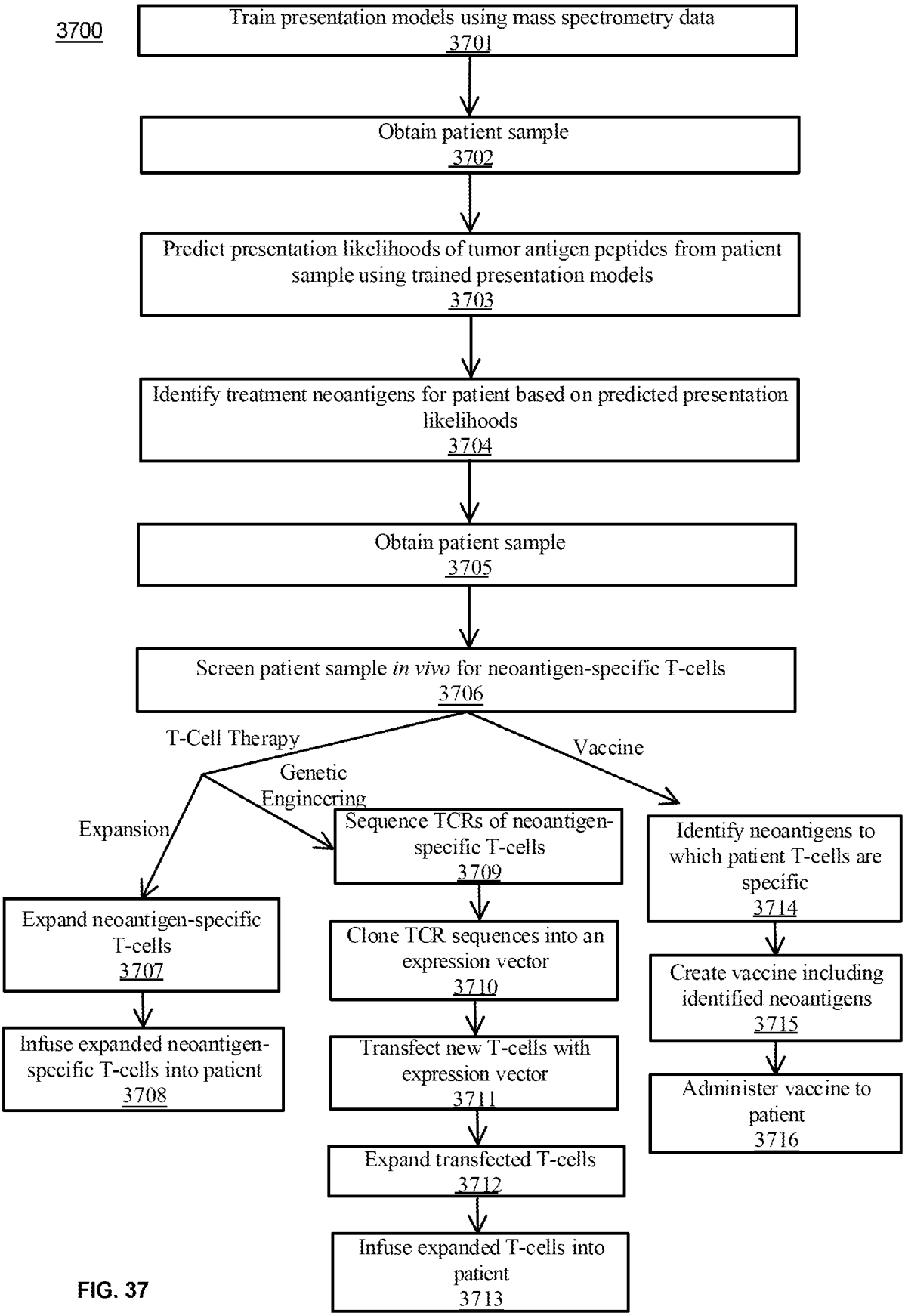

3700
Train presentation models using mass spectrometry data
3701
Obtain patient sample
3702
Predict presentation likelihoods of tumor antigen peptides from patient sample using trained presentation models
3703
Identify treatment neoantigens for patient based on predicted presentation likelihoods
3704
Obtain patient sample
3705
Screen patient sample *in vivo* for neoantigen-specific T-cells
3706
T-Cell Therapy
Vaccine
Genetic Engineering
Expansion
Sequence TCRs of neoantigen-specific T-cells
3709
Identify neoantigens to which patient T-cells are specific
3714
Expand neoantigen-specific T-cells
3707
Clone TCR sequences into an expression vector
3710
Create vaccine including identified neoantigens
3715
Infuse expanded neoantigen-specific T-cells into patient
3708
Transfect new T-cells with expression vector
3711
Administer vaccine to patient
3716
Expand transfected T-cells
3712
Infuse expanded T-cells into patient
3713

FIG. 37

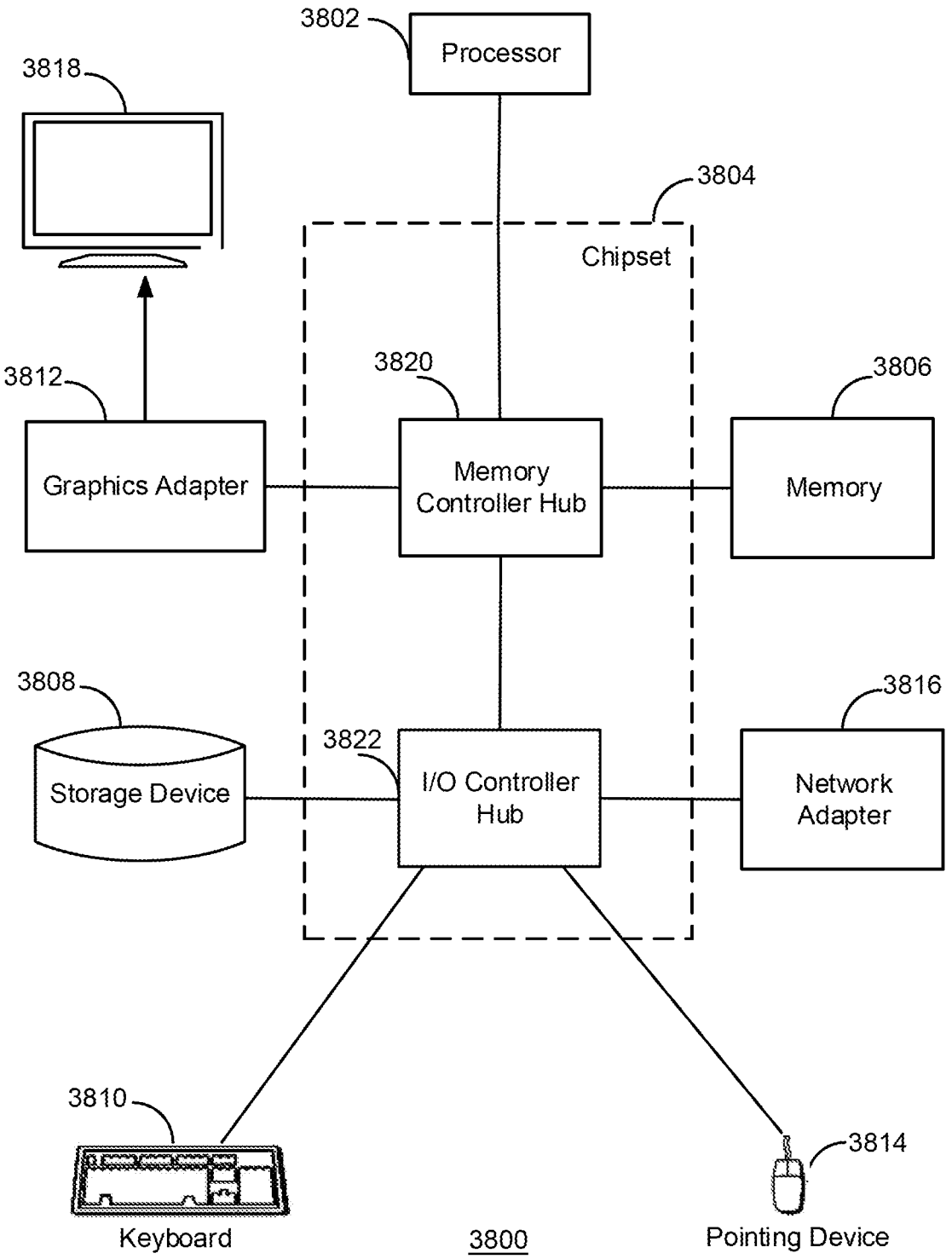

3802
Processor
3818
3804
Chipset
3812
3820
3806
Graphics Adapter
Memory Controller Hub
Memory
3808
3816
3822
Storage Device
I/O Controller Hub
Network Adapter
3810
3814
Keyboard
3800
Pointing Device

FIG. 38

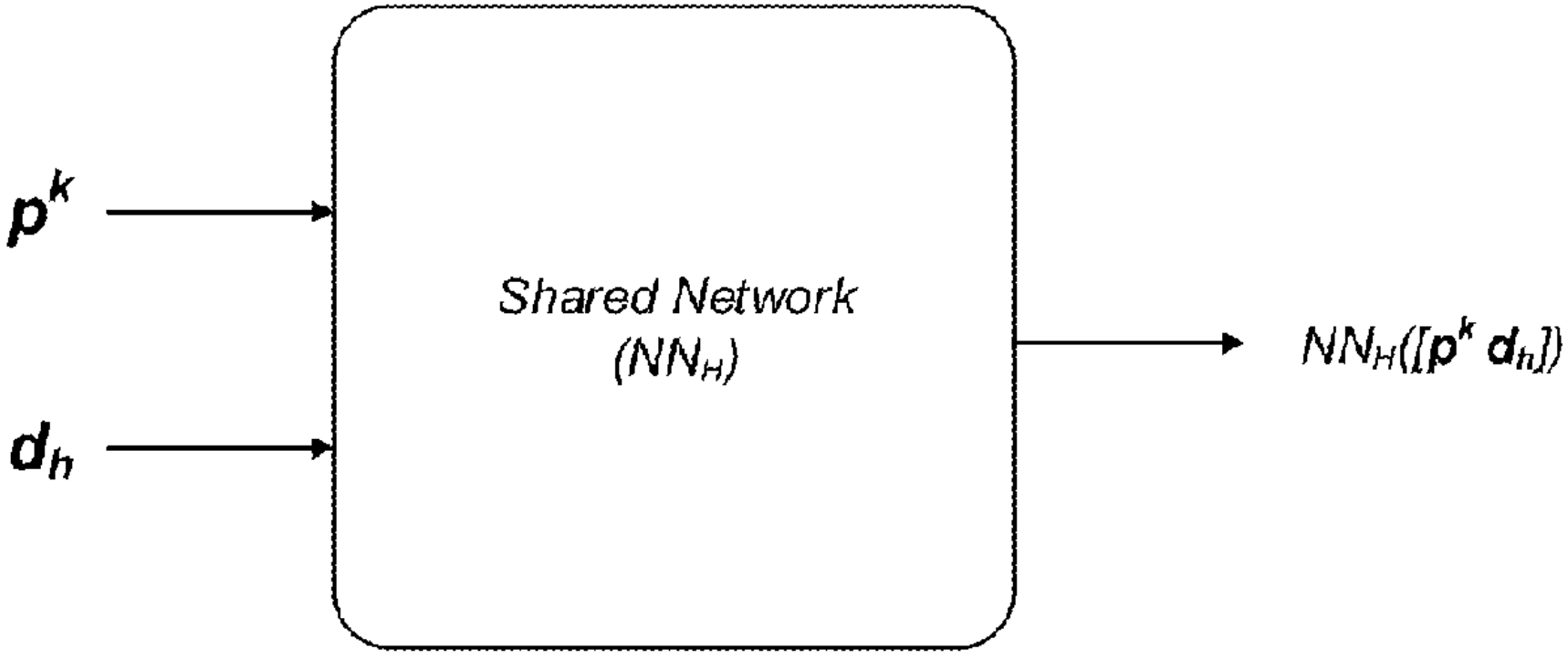


FIG. 13