



## (51) International Patent Classification:

G06F 19/22 (2011.01) A61K 35/15 (2015.01)  
A61K 39/00 (2006.01) C07K 7/08 (2006.01)  
A61K 38/10 (2006.01)

## (21) International Application Number:

PCT/US2018/036571

## (22) International Filing Date:

08 June 2018 (08.06.2018)

## (25) Filing Language:

English

## (26) Publication Language:

English

## (30) Priority Data:

62/517,786 09 June 2017 (09.06.2017) US

(71) Applicant: GRITSTONE ONCOLOGY, INC. [US/US];  
5858 Horton Street, Suite 210, Emeryville, CA 94608 (US).

(72) Inventors: **BULIK-SULLIVAN, Brendan**; Gritstone Oncology, Inc., 5858 Horton Street, Suite 210, Emeryville, CA 94608 (US). **YELENSKY, Roman**; Gritstone Oncology, Inc., 5858 Horton Street, Suite 210, Emeryville, CA 94608 (US). **BUSBY, Jennifer**; Gritstone Oncology, Inc., 5858 Horton Street, Suite 210, Emeryville, CA 94608 (US). **DETTI, Adnan**; Gritstone Oncology, Inc., 5858 Horton Street, Suite 210, Emeryville, CA 94608 (US).

(74) Agent: **WHITEHEAD, Andrew, P.** et al.; Fenwick & West LLP, 801 California Street, Mountain View, CA 94041 (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,

## (54) Title: NEOANTIGEN IDENTIFICATION, MANUFACTURE, AND USE

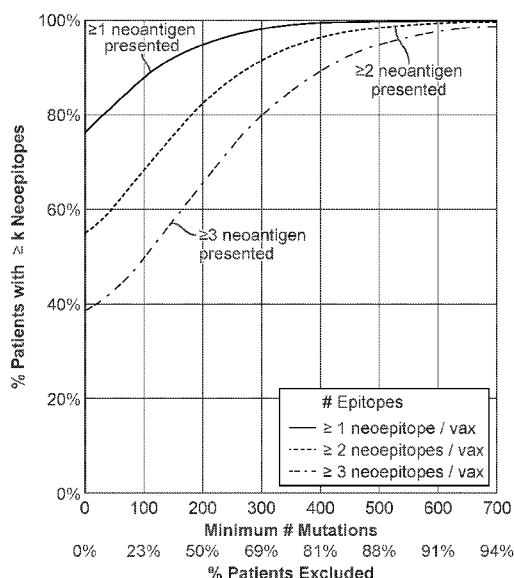


FIG. 13B

(57) Abstract: Disclosed herein is a system and methods for determining the alleles, neoantigens, and vaccine composition as determined on the basis of an individual's tumor mutations. Also disclosed are systems and methods for obtaining high quality sequencing data from a tumor. Further, described herein are systems and methods for identifying somatic changes in polymorphic genome data. Further, described herein are systems and methods for selecting a subset of patients for treatment. A utility score indicating an estimated number of neoantigens presented on the surface of tumor cells is determined for each patient based on one or more neoantigen candidates identified for the patient. The subset of patients are selected based on the determined utility scores. The selected subset of patients can receive treatment, such as neoantigen vaccines or checkpoint inhibitor therapy. Finally, described herein are unique cancer vaccines.



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, 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).

**Declarations under Rule 4.17:**

— *of inventorship (Rule 4.17(iv))*

**Published:**

— *with international search report (Art. 21(3))*  
 — *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

**TITLE**

Neoantigen identification, manufacture, and use.

**CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims the benefit of U.S. Provisional Application No. 62/517,786, filed June 9, 2017, which is hereby incorporated by reference in its entirety.

**BACKGROUND**

[0002] Therapeutic vaccines based on tumor-specific neoantigens hold great promise as a next-generation of personalized cancer immunotherapy.<sup>1-3</sup> 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.<sup>4,5</sup> Early evidence shows that neoantigen-based vaccination can elicit T-cell responses<sup>6</sup> and that neoantigen targeted cell-therapy can cause tumor regression under certain circumstances in selected patients.<sup>7</sup> Both MHC class I and MHC class II have an impact on T-cell responses<sup>70-71</sup>.

[0003] One question for neoantigen vaccine design is which of the many coding mutations present in subject tumors can generate the “best” therapeutic neoantigens, e.g., antigens that can elicit anti-tumor immunity and cause tumor regression.

[0004] 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<sup>8</sup>. 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<sup>9</sup>. Consequently, existing methods are likely to suffer from reduced low positive predictive value (PPV). (Figure 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<sup>10,11</sup> (Figure

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.<sup>12</sup>

**[0006]** This low positive predictive value (PPV) of existing methods for predicting presentation presents a problem for neoantigen-based vaccine design. 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). Thus, neoantigen vaccination with current methods is unlikely to succeed in a substantial number of subjects having tumors. (Figure 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<sup>13</sup>, 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. 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 are more likely to elicit anti-tumor immunity. These approaches include, depending on the embodiment, trained statistic regression or nonlinear deep learning models that jointly model peptide-allele mappings as

well as the per-allele motifs for peptide of multiple lengths, sharing statistical strength across peptides of different lengths. 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 are addressed..

**[0010]** Also disclosed herein is a method of identifying a subset of patients for treatment. At least one of exome, transcriptome, or whole genome tumor nucleotide sequencing data from tumor cells and normal cells of each patient are obtained. The tumor nucleotide sequencing data is used to obtain 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. The peptide sequence of each neoantigen for the patient includes at least one alteration that makes it distinct from a corresponding wild-type parental peptide sequence identified from the normal cells of the patient. A set of numerical presentation likelihoods for the set of neoantigens for each patient are generated by inputting the peptide sequences of each of the set of neoantigens into a machine-learned presentation model. Each presentation likelihood represents the likelihood that a corresponding neoantigen is presented by one or more MHC alleles on the surface of the tumor cells of the patient. The set of presentation likelihoods have been identified at least based on mass spectrometry data. One or more neoantigens from the set of neoantigens for the patient are identified. A utility score indicating an estimated number of neoantigens presented on the surface of the tumor cells of the patient as determined by the corresponding presentation likelihoods for the one or more neoantigens for the patient is determined for each patient. A subset of patients are selected for treatment. Each patient in the subset of patients is associated with a utility score that satisfies a predetermined inclusion criteria. The selected subset of patients can receive treatment, such as neoantigen vaccines or checkpoint inhibitor therapy.

#### **BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS**

**[0011]** 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:

**[0012]** Figure (FIG.) 1A shows current clinical approaches to neoantigen identification.

- [0013] FIG. 1B shows that <5% of predicted bound peptides are presented on tumor cells.
- [0014] FIG. 1C shows the impact of the neoantigen prediction specificity problem.
- [0015] FIG. 1D shows that binding prediction is not sufficient for neoantigen identification.
- [0016] FIG. 1E shows probability of MHC-I presentation as a function of peptide length.
- [0017] FIG. 1F shows an example peptide spectrum generated from Promega's dynamic range standard.
- [0018] FIG. 1G shows how the addition of features increases the model positive predictive value.
- [0019] FIG. 2A is an overview of an environment for identifying likelihoods of peptide presentation in patients, in accordance with an embodiment.
- [0020] FIG. 2B and 2C illustrate a method of obtaining presentation information, in accordance with an embodiment.
- [0021] FIG. 3 is a high-level block diagram illustrating the computer logic components of the presentation identification system, according to one embodiment.
- [0022] FIG. 4 illustrates an example set of training data, according to one embodiment.
- [0023] FIG. 5 illustrates an example network model in association with an MHC allele. FIG. 6A illustrates an example network model  $NN_H(\cdot)$  shared by MHC alleles, according to one embodiment.
- [0024] FIG. 6B illustrates an example network model  $NN_H(\cdot)$  shared by MHC alleles, according to another embodiment.
- [0025] FIG. 7 illustrates generating a presentation likelihood for a peptide in association with an MHC allele using an example network model.
- [0026] FIG. 8 illustrates generating a presentation likelihood for a peptide in association with a MHC allele using example network models.
- [0027] FIG. 9 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.
- [0028] FIG. 10 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.
- [0029] FIG. 11 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.
- [0030] FIG. 12 illustrates generating a presentation likelihood for a peptide in association with MHC alleles using example network models.

[0031] FIG. 13A illustrates a sample frequency distribution of tumor mutation burden in NSCLC patients.

[0032] FIG. 13B illustrates the number of presented neoantigens in simulated vaccines for patients selected based on an inclusion criteria of whether the patients satisfy a minimum tumor mutation burden, in accordance with an embodiment.

[0033] FIG. 13C 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.

[0034] FIG. 13D 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.

[0035] FIG. 13E compares the number of presented neoantigens in simulated vaccines between patients selected based on tumor mutation burden and patients selected by expectation utility score, in accordance with an embodiment.

[0036] FIG. 14 illustrates an example computer for implementing the entities shown in FIGS. 1 and 3.

## DETAILED DESCRIPTION

### I. Definitions

[0037] 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.

[0038] As used herein the term “antigen” is a substance that induces an immune response.

[0039] 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.

**[0040]** 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.

**[0041]** As used herein the term “neoantigen-based vaccine” is a vaccine construct based on one or more neoantigens, e.g., a plurality of neoantigens.

**[0042]** As used herein the term “candidate neoantigen” is a mutation or other aberration giving rise to a new sequence that may represent a neoantigen.

**[0043]** As used herein the term “coding region” is the portion(s) of a gene that encode protein.

**[0044]** As used herein the term “coding mutation” is a mutation occurring in a coding region.

**[0045]** As used herein the term “ORF” means open reading frame.

**[0046]** As used herein the term “NEO-ORF” is a tumor-specific ORF arising from a mutation or other aberration such as splicing.

**[0047]** As used herein the term “missense mutation” is a mutation causing a substitution from one amino acid to another.

**[0048]** As used herein the term “nonsense mutation” is a mutation causing a substitution from an amino acid to a stop codon.

**[0049]** As used herein the term “frameshift mutation” is a mutation causing a change in the frame of the protein.

**[0050]** As used herein the term “indel” is an insertion or deletion of one or more nucleic acids.

**[0051]** 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.

**[0052]** 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).

**[0053]** 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*).

**[0054]** 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.

**[0055]** As used herein the term "non-stop or read-through" is a mutation causing the removal of the natural stop codon.

**[0056]** As used herein the term "epitope" is the specific portion of an antigen typically bound by an antibody or T cell receptor.

**[0057]** As used herein the term "immunogenic" is the ability to elicit an immune response, e.g., via T cells, B cells, or both.

**[0058]** As used herein the term "HLA binding affinity" "MHC binding affinity" means affinity of binding between a specific antigen and a specific MHC allele.

- [0059] As used herein the term “bait” is a nucleic acid probe used to enrich a specific sequence of DNA or RNA from a sample.
- [0060] As used herein the term “variant” is a difference between a subject’s nucleic acids and the reference human genome used as a control.
- [0061] As used herein the term “variant call” is an algorithmic determination of the presence of a variant, typically from sequencing.
- [0062] As used herein the term “polymorphism” is a germline variant, i.e., a variant found in all DNA-bearing cells of an individual.
- [0063] As used herein the term “somatic variant” is a variant arising in non-germline cells of an individual.
- [0064] 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.
- [0065] As used herein the term “HLA type” is the complement of HLA gene alleles.
- [0066] 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.
- [0067] 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.
- [0068] 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.
- [0069] 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.
- [0070] 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.
- [0071] 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.
- [0072] As used herein the term “proteome” is the set of all proteins expressed and/or translated by a cell, group of cells, or individual.
- [0073] 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).

**[0074]** 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.

**[0075]** As used herein the term “dextramers” is a dextran-based peptide-MHC multimers used for antigen-specific T-cell staining in flow cytometry.

**[0076]** 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.

**[0077]** 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).

**[0078]** 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.

**[0079]** 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.

**[0080]** 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.

**[0081]** 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.

**[0082]** 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.

[0083] 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.

[0084] 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.

[0085] 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.

[0086] 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**

[0087] Disclosed herein is are methods for identifying neoantigens from a tumor of a subject that are likely to be presented on the cell surface of the tumor and/or are likely to be immunogenic. As an example, one such method may comprise the steps of: obtaining at least one of exome, transcriptome or whole genome tumor nucleotide sequencing data from the tumor cell of the subject, wherein the tumor nucleotide sequencing data is used to obtain data representing peptide sequences of each of a set of neoantigens, and wherein the peptide sequence of each neoantigen comprises at least one alteration that makes it distinct from the corresponding wild-type, parental peptide sequence; inputting the peptide sequence of each neoantigen into one or more presentation models to generate a set of numerical likelihoods that each of the neoantigens is presented by one or more MHC alleles on the tumor cell

surface of the tumor cell of the subject or cells present in the tumor, the set of numerical likelihoods having been identified at least based on received mass spectrometry data; and selecting a subset of the set of neoantigens based on the set of numerical likelihoods to generate a set of selected neoantigens.

**[0088]** The presentation model can comprise a statistical regression or a machine learning (e.g., deep learning) model trained on a set of reference data (also referred to as a training data set) comprising a set of corresponding labels, wherein the set of reference data is obtained from each of a plurality of distinct subjects where optionally some subjects can have a tumor, and wherein the set of reference data comprises at least one of: data representing exome nucleotide sequences from tumor tissue, data representing exome nucleotide sequences from normal tissue, data representing transcriptome nucleotide sequences from tumor tissue, data representing proteome sequences from tumor tissue, and data representing MHC peptidome sequences from tumor tissue, and data representing MHC peptidome sequences from normal tissue. The reference data can further comprise mass spectrometry data, sequencing data, RNA sequencing data, and proteomics data for single-allele cell lines engineered to express a predetermined MHC allele that are subsequently exposed to synthetic protein, normal and tumor human cell lines, and fresh and frozen primary samples, and T cell assays (e.g., ELISPOT). In certain aspects, the set of reference data includes each form of reference data.

**[0089]** The presentation model can comprise a set of features derived at least in part from the set of reference data, and wherein the set of features comprises at least one of allele dependent-features and allele-independent features. In certain aspects each feature is included.

**[0090]** Dendritic cell presentation to naïve T cell features can comprise at least one of: A feature described above. The dose and type of antigen in the vaccine. (e.g., peptide, mRNA, virus, etc.): (1) The route by which dendritic cells (DCs) take up the antigen type (e.g., endocytosis, micropinocytosis); and/or (2) The efficacy with which the antigen is taken up by DCs. The dose and type of adjuvant in the vaccine. The length of the vaccine antigen sequence. The number and sites of vaccine administration. Baseline patient immune functioning (e.g., as measured by history of recent infections, blood counts, etc). For RNA vaccines: (1) the turnover rate of the mRNA protein product in the dendritic cell; (2) the rate of translation of the mRNA after uptake by dendritic cells as measured in in vitro or in vivo experiments; and/or (3) the number or rounds of translation of the mRNA after uptake by

dendritic cells as measured by in vivo or in vitro experiments. The presence of protease cleavage motifs in the peptide, optionally giving additional weight to proteases typically expressed in dendritic cells (as measured by RNA-seq or mass spectrometry). The level of expression of the proteasome and immunoproteasome in typical activated dendritic cells (which may be measured by RNA-seq, mass spectrometry, immunohistochemistry, or other standard techniques). The expression levels of the particular MHC allele in the individual in question (e.g., as measured by RNA-seq or mass spectrometry), optionally measured specifically in activated dendritic cells or other immune cells. The probability of peptide presentation by the particular MHC allele in other individuals who express the particular MHC allele, optionally measured specifically in activated dendritic cells or other immune cells. The probability of peptide 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, optionally measured specifically in activated dendritic cells or other immune cells.

**[0091]** Immune tolerance escape features can comprise at least one of: Direct measurement of the self-peptidome via protein mass spectrometry performed on one or several cell types. Estimation of the self-peptidome by taking the union of all k-mer (e.g. 5-25) substrings of self-proteins. Estimation of the self-peptidome using a model of presentation similar to the presentation model described above applied to all non-mutation self-proteins, optionally accounting for germline variants.

**[0092]** Ranking can be performed using the plurality of neoantigens provided by at least one model based at least in part on the numerical likelihoods. Following the ranking a selecting can be performed to select a subset of the ranked neoantigens according to a selection criteria. After selecting a subset of the ranked peptides can be provided as an output.

**[0093]** A number of the set of selected neoantigens may be 20.

**[0094]** The presentation model may represent dependence between presence of a pair of a particular one of the MHC alleles and a particular amino acid at a particular position of a peptide sequence; and likelihood of presentation on the tumor cell surface, by the particular one of the MHC alleles of the pair, of such a peptide sequence comprising the particular amino acid at the particular position.

**[0095]** A method disclosed herein can also include applying the one or more presentation models to the peptide sequence of the corresponding neoantigen to generate a dependency score for each of the one or more MHC alleles indicating whether the MHC allele will

present the corresponding neoantigen based on at least positions of amino acids of the peptide sequence of the corresponding neoantigen.

**[0096]** A method disclosed herein can also include 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 numerical likelihood.

**[0097]** The step of transforming the dependency scores can model the presentation of the peptide sequence of the corresponding neoantigen as mutually exclusive.

**[0098]** A method disclosed herein can also include transforming a combination of the dependency scores to generate the numerical likelihood.

**[0099]** The step of transforming the combination of the dependency scores can model the presentation of the peptide sequence of the corresponding neoantigen as interfering between MHC alleles.

**[00100]** The set of numerical likelihoods can be further identified by at least an allele noninteracting feature, and a method disclosed herein can also include applying an allele noninteracting one of the one or more presentation models 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.

**[00101]** A method disclosed herein can also include combining the dependency score for each MHC allele in the one or more MHC alleles with the dependency score for the allele noninteracting feature; transforming the combined dependency scores for each MHC allele to generate a corresponding per-allele likelihood for the MHC allele indicating a likelihood that the corresponding MHC allele will present the corresponding neoantigen; and combining the per-allele likelihoods to generate the numerical likelihood.

**[00102]** A method disclosed herein can also include transforming a combination of the dependency scores for each of the MHC alleles and the dependency score for the allele noninteracting features to generate the numerical likelihood.

**[00103]** A set of numerical parameters for the presentation model can be trained based on a training data set including at least a set of training peptide sequences identified as present in a plurality of samples and one or more MHC alleles associated with each training peptide sequence, wherein the training peptide sequences are identified through mass spectrometry on isolated peptides eluted from MHC alleles derived from the plurality of samples.

**[00104]** The samples can also include cell lines engineered to express a single MHC class I or class II allele.

**[00105]** The samples can also include cell lines engineered to express a plurality of MHC class I or class II alleles.

**[00106]** The samples can also include human cell lines obtained or derived from a plurality of patients.

**[00107]** The samples can also include fresh or frozen tumor samples obtained from a plurality of patients.

**[00108]** The samples can also include fresh or frozen tissue samples obtained from a plurality of patients.

**[00109]** The samples can also include peptides identified using T-cell assays.

**[00110]** The training data set can further include data associated with: peptide abundance of the set of training peptides present in the samples; peptide length of the set of training peptides in the samples.

**[00111]** The training data set may be generated by comparing the set of training peptide sequences via alignment to a database comprising a set of known protein sequences, wherein the set of training protein sequences are longer than and include the training peptide sequences.

**[00112]** The training data set may be generated based on performing or having performed nucleotide sequencing on a cell line to obtain at least one of exome, transcriptome, or whole genome sequencing data from the cell line, the sequencing data including at least one nucleotide sequence including an alteration.

**[00113]** The training data set may be generated based on obtaining at least one of exome, transcriptome, and whole genome normal nucleotide sequencing data from normal tissue samples.

**[00114]** The training data set may further include data associated with proteome sequences associated with the samples.

**[00115]** The training data set may further include data associated with MHC peptidome sequences associated with the samples.

**[00116]** The training data set may further include data associated with peptide-MHC binding affinity measurements for at least one of the isolated peptides.

**[00117]** The training data set may further include data associated with peptide-MHC binding stability measurements for at least one of the isolated peptides.

**[00118]** The training data set may further include data associated with transcriptomes associated with the samples.

**[00119]** The training data set may further include data associated with genomes associated with the samples.

**[00120]** The training peptide sequences may be of lengths within a range of k-mers where k is between 8-15, inclusive for MHC class I or 6-30 inclusive for MHC class II.

**[00121]** A method disclosed herein can also include encoding the peptide sequence using a one-hot encoding scheme.

**[00122]** A method disclosed herein can also include encoding the training peptide sequences using a left-padded one-hot encoding scheme.

**[00123]** A method of treating a subject having a tumor, comprising performing the steps of claim 1, and further comprising obtaining a tumor vaccine comprising the set of selected neoantigens, and administering the tumor vaccine to the subject.

**[00124]** Also disclosed herein is a methods for manufacturing a tumor vaccine, comprising the steps of: obtaining at least one of exome, transcriptome or whole genome tumor nucleotide sequencing data from the tumor cell of the subject, wherein the tumor nucleotide sequencing data is used to obtain data representing peptide sequences of each of a set of neoantigens, and wherein the peptide sequence of each neoantigen comprises at least one mutation that makes it distinct from the corresponding wild-type, parental peptide sequence; inputting the peptide sequence of each neoantigen into one or more presentation models to generate a set of numerical likelihoods that each of the neoantigens is presented by one or more MHC alleles on the tumor cell surface of the tumor cell of the subject, the set of numerical likelihoods having been identified at least based on received mass spectrometry data; and selecting a subset of the set of neoantigens based on the set of numerical likelihoods to generate a set of selected neoantigens; and producing or having produced a tumor vaccine comprising the set of selected neoantigens.

**[00125]** Also disclosed herein is a tumor vaccine including a set of selected neoantigens selected by performing the method comprising the steps of: obtaining at least one of exome, transcriptome or whole genome tumor nucleotide sequencing data from the tumor cell of the subject, wherein the tumor nucleotide sequencing data is used to obtain data representing peptide sequences of each of a set of neoantigens, and wherein the peptide sequence of each neoantigen comprises at least one mutation that makes it distinct from the corresponding wild-type, parental peptide sequence; inputting the peptide sequence of each neoantigen into

one or more presentation models to generate a set of numerical likelihoods that each of the neoantigens is presented by one or more MHC alleles on the tumor cell surface of the tumor cell of the subject, the set of numerical likelihoods having been identified at least based on received mass spectrometry data; and selecting a subset of the set of neoantigens based on the set of numerical likelihoods to generate a set of selected neoantigens; and producing or having produced a tumor vaccine comprising the set of selected neoantigens.

**[00126]** The tumor vaccine may include one or more of a nucleotide sequence, a polypeptide sequence, RNA, DNA, a cell, a plasmid, or a vector.

**[00127]** The tumor vaccine may include one or more neoantigens presented on the tumor cell surface.

**[00128]** The tumor vaccine may include one or more neoantigens that is immunogenic in the subject.

**[00129]** The tumor vaccine may not include one or more neoantigens that induce an autoimmune response against normal tissue in the subject.

**[00130]** The tumor vaccine may include an adjuvant.

**[00131]** The tumor vaccine may include an excipient.

**[00132]** A method disclosed herein may also include selecting neoantigens that have an increased likelihood of being presented on the tumor cell surface relative to unselected neoantigens based on the presentation model.

**[00133]** A method disclosed herein may also include 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 presentation model.

**[00134]** A method disclosed herein may also include 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).

**[00135]** A method disclosed herein may also include selecting neoantigens that have a decreased likelihood of being subject to inhibition via central or peripheral tolerance relative to unselected neoantigens based on the presentation model.

**[00136]** A method disclosed herein may also include 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 presentation model.

**[00137]** The exome or transcriptome nucleotide sequencing data may be obtained by performing sequencing on the tumor tissue.

**[00138]** The sequencing may be next generation sequencing (NGS) or any massively parallel sequencing approach.

**[00139]** The set of numerical likelihoods may be further identified by at least MHC-allele interacting features comprising at least one of: the predicted affinity with which the MHC allele and the neoantigen encoded peptide bind; the predicted stability of the neoantigen encoded peptide-MHC complex; the sequence and length of the neoantigen encoded peptide; the probability of presentation of neoantigen encoded peptides with similar sequence in cells from other individuals expressing the particular MHC allele as assessed by mass-spectrometry proteomics or other means; the expression levels of the particular MHC allele in the subject in question (e.g. as measured by RNA-seq or mass spectrometry); the overall neoantigen encoded peptide-sequence-independent probability of presentation by the particular MHC allele in other distinct subjects who express the particular MHC allele; the overall neoantigen encoded 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 distinct subjects.

**[00140]** The set of numerical likelihoods are further identified by at least MHC-allele noninteracting features comprising at least one of: the C- and N-terminal sequences flanking the neoantigen encoded peptide within its source protein sequence; the presence of protease cleavage motifs in the neoantigen encoded peptide, optionally weighted according to the expression of corresponding proteases in the tumor cells (as measured by RNA-seq or mass spectrometry); the turnover rate of the source protein as measured in the appropriate cell type; 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; 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); the expression of the source gene of the neoantigen encoded peptide (e.g., as measured by RNA-seq or mass spectrometry); the typical tissue-specific expression of the source gene of the neoantigen encoded peptide during various stages of the cell cycle; a comprehensive catalog of features of the source protein and/or its domains as can be found in e.g. uniProt or PDB

<http://www.rcsb.org/pdb/home/home.do>; 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; the probability of presentation of peptides from the source protein of the neoantigen encoded peptide in question in other distinct subjects; the probability that the peptide will not be detected or over-represented by mass spectrometry due to technical biases; the expression of various gene modules/pathways as measured by RNASeq (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); the copy number of the source gene of the neoantigen encoded peptide in the tumor cells; the probability that the peptide binds to the TAP or the measured or predicted binding affinity of the peptide to the TAP; the expression level of TAP in the tumor cells (which may be measured by RNA-seq, proteome mass spectrometry, immunohistochemistry); presence or absence of tumor mutations, including, but not limited to: driver mutations in known cancer driver genes such as EGFR, KRAS, ALK, RET, ROS1, TP53, CDKN2A, CDKN2B, NTRK1, NTRK2, NTRK3, and 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-DOB, HLA-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; presence or absence of functional germline polymorphisms, including, but not limited to: 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-DOB, HLA-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); tumor type (e.g., NSCLC, melanoma); clinical tumor subtype (e.g., squamous lung cancer vs. non-squamous); smoking history; the typical expression of the source gene of the peptide in the relevant tumor type or clinical subtype, optionally stratified by driver mutation.

**[00141]** The at least one mutation may be 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.

**[00142]** The tumor cell may be 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.

**[00143]** A method disclosed herein may also include obtaining a tumor vaccine comprising the set of selected neoantigens or a subset thereof, optionally further comprising administering the tumor vaccine to the subject.

**[00144]** At least one of neoantigens in the set of selected neoantigens, when in polypeptide form, may include at least one of: a binding affinity with MHC with an IC<sub>50</sub> value of less than 1000nM, for MHC Class I polypeptides a length of 8-15, 8, 9, 10, 11, 12, 13, 14, or 15 amino acids, for MHC Class II polypeptides a length of 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 polypeptide in the parent protein sequence promoting proteasome cleavage, and presence of sequence motifs promoting TAP transport. For MHC Class II, 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.

**[00145]** Also disclosed herein is a methods for generating a model for identifying one or more neoantigens that are likely to be presented on a tumor cell surface of a tumor cell, comprising the steps of: receiving mass spectrometry data comprising data associated with a plurality of isolated peptides eluted from major histocompatibility complex (MHC) derived from a plurality of samples; obtaining a training data set by at least identifying a set of training peptide sequences present in the samples and one or more MHCs associated with each training peptide sequence; training a set of numerical parameters of a presentation model using the training data set comprising the training peptide sequences, the presentation model providing a plurality of numerical likelihoods that peptide sequences from the tumor cell are presented by one or more MHC alleles on the tumor cell surface.

**[00146]** The presentation model may represent dependence between: presence of a particular amino acid at a particular position of a peptide sequence; and likelihood of

presentation, by one of the MHC alleles on the tumor cell, of the peptide sequence containing the particular amino acid at the particular position.

**[00147]** The samples can also include cell lines engineered to express a single MHC class I or class II allele.

**[00148]** The samples can also include cell lines engineered to express a plurality of MHC class I or class II alleles.

**[00149]** The samples can also include human cell lines obtained or derived from a plurality of patients.

**[00150]** The samples can also include fresh or frozen tumor samples obtained from a plurality of patients.

**[00151]** The samples can also include peptides identified using T-cell assays.

**[00152]** The training data set may further include data associated with: peptide abundance of the set of training peptides present in the samples; peptide length of the set of training peptides in the samples.

**[00153]** A method disclosed herein can also include obtaining a set of training protein sequences based on the training peptide sequences by comparing the set of training peptide sequences via alignment to a database comprising a set of known protein sequences, wherein the set of training protein sequences are longer than and include the training peptide sequences.

**[00154]** A method disclosed herein can also include performing or having performed mass spectrometry on a cell line to obtain at least one of exome, transcriptome, or whole genome nucleotide sequencing data from the cell line, the nucleotide sequencing data including at least one protein sequence including a mutation.

**[00155]** A method disclosed herein can also include: encoding the training peptide sequences using a one-hot encoding scheme.

**[00156]** A method disclosed herein can also include obtaining at least one of exome, transcriptome, and whole genome normal nucleotide sequencing data from normal tissue samples; and training the set of parameters of the presentation model using the normal nucleotide sequencing data.

**[00157]** The training data set may further include data associated with proteome sequences associated with the samples.

**[00158]** The training data set may further include data associated with MHC peptidome sequences associated with the samples.

- [00159] The training data set may further include data associated with peptide-MHC binding affinity measurements for at least one of the isolated peptides.
- [00160] The training data set may further include data associated with peptide-MHC binding stability measurements for at least one of the isolated peptides.
- [00161] The training data set may further include data associated with transcriptomes associated with the samples.
- [00162] The training data set may further include data associated with genomes associated with the samples.
- [00163] A method disclosed herein may also include logistically regressing the set of parameters.
- [00164] The training peptide sequences may be lengths within a range of k-mers where k is between 8-15, inclusive for MHC class I or 6-30, inclusive for MHC class II.
- [00165] A method disclosed herein may also include encoding the training peptide sequences using a left-padded one-hot encoding scheme.
- [00166] A method disclosed herein may also include determining values for the set of parameters using a deep learning algorithm.
- [00167] Disclosed herein is are methods for identifying one or more neoantigens that are likely to be presented on a tumor cell surface of a tumor cell, comprising executing the steps of: receiving mass spectrometry data comprising data associated with a plurality of isolated peptides eluted from major histocompatibility complex (MHC) derived from a plurality of fresh or frozen tumor samples; obtaining a training data set by at least identifying a set of training peptide sequences present in the tumor samples and presented on one or more MHC alleles associated with each training peptide sequence; obtaining a set of training protein sequences based on the training peptide sequences; and training a set of numerical parameters of a presentation model using the training protein sequences and the training peptide sequences, the presentation model providing a plurality of numerical likelihoods that peptide sequences from the tumor cell are presented by one or more MHC alleles on the tumor cell surface.
- [00168] The presentation model may represent dependence between: presence of a pair of a particular one of the MHC alleles and a particular amino acid at a particular position of a peptide sequence; and likelihood of presentation on the tumor cell surface, by the particular one of the MHC alleles of the pair, of such a peptide sequence comprising the particular amino acid at the particular position.

**[00169]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has an increased likelihood that it is presented on the cell surface of the tumor relative to one or more distinct tumor neoantigens.

**[00170]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has an increased likelihood that it is capable of inducing a tumor-specific immune response in the subject relative to one or more distinct tumor neoantigens.

**[00171]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has an increased likelihood that it is capable of being presented to naïve T cells by professional antigen presenting cells (APCs) relative to one or more distinct tumor neoantigens, optionally wherein the APC is a dendritic cell (DC).

**[00172]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has a decreased likelihood that it is subject to inhibition via central or peripheral tolerance relative to one or more distinct tumor neoantigens.

**[00173]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has a decreased likelihood that it is capable of inducing an autoimmune response to normal tissue in the subject relative to one or more distinct tumor neoantigens.

**[00174]** A method disclosed herein can also include selecting a subset of neoantigens, wherein the subset of neoantigens is selected because each has a decreased likelihood that it will be differentially post-translationally modified in tumor cells versus APCs, optionally wherein the APC is a dendritic cell (DC).

**[00175]** The practice of the methods herein will employ, unless otherwise indicated, conventional methods of protein chemistry, biochemistry, recombinant DNA techniques and pharmacology, within the skill of the art. Such techniques are explained fully in the literature. See, e.g., T.E. Creighton, *Proteins: Structures and Molecular Properties* (W.H. Freeman and Company, 1993); A.L. Lehninger, *Biochemistry* (Worth Publishers, Inc., current addition); Sambrook, et al., *Molecular Cloning: A Laboratory Manual* (2nd Edition, 1989); *Methods In Enzymology* (S. Colowick and N. Kaplan eds., Academic Press, Inc.); *Remington's Pharmaceutical Sciences*, 18th Edition (Easton, Pennsylvania: Mack Publishing

Company, 1990); Carey and Sundberg *Advanced Organic Chemistry 3<sup>rd</sup> Ed.* (Plenum Press) Vols A and B(1992).

**[00176]** The set of presentation likelihoods can also be generated based on source genes of the set of neoantigens.

**[00177]** The set of presentation likelihoods can also be generated based on source genes and source tissue type of the set of neoantigens.

**[00178]** A method disclosed herein can also include identifying a subset of patients for treatment with neoantigen vaccines, the steps comprising: obtaining, for each patient, at least one of exome, transcriptome, or whole genome tumor nucleotide sequencing data from a tumor cell of the patient, the tumor nucleotide sequencing data used to obtain peptide sequences of each of a set of neoantigens, and the peptide sequence of each neoantigen comprising at least one alteration that makes it distinct from a corresponding wild-type, parental peptide sequence; generating, for each patient, a set of numerical presentation likelihoods for the set of neoantigens for the patient by inputting the peptide sequences of each of the set of neoantigens into one or more presentation models, the set of presentation likelihoods indicating likelihood that each of the set of neoantigens is presented by one or more MHC alleles on a surface of the tumor cell of the patient, the set of presentation likelihoods having been identified at least based on received mass spectrometry data; identifying, for each patient, a treatment subset of neoantigens from the set of neoantigens of the patient, the treatment subset corresponding to a predetermined number of neoantigens having highest presentation likelihoods in the set of presentation likelihoods generated for the patient; and selecting the subset of patients for treatment with neoantigen vaccines, the selected subset of patients satisfying an inclusion criteria based on the set of neoantigens obtained for each patient in the selected subset or based on the tumor nucleotide sequencing data.

**[00179]** A method disclosed herein may also include treating each patient in the selected subset of patients with a corresponding neoantigen vaccine, the neoantigen vaccine for the patient including the treatment subset identified through the set of presentation likelihoods for the patient.

**[00180]** A method disclosed herein may also include selecting the subset of patients having tumor mutation burden (TMB) above a minimum threshold, the TMB for a patient indicating a number of neoantigens in the set of neoantigens associated with the patient.

[00181] A method disclosed herein may also include identifying, for each patient, a utility score indicating a measure of an estimated number of presented neoantigens from the treatment subset of the patient; and selecting the subset of patients having utility scores above a minimum threshold.

[00182] The neoantigen presentation may be modeled as a Bernoulli random variable, the utility score can represent an expected number of presented neoantigens in the treatment subset for the patient, the utility score can be given by a summation of the presentation likelihoods for each neoantigen in the treatment subset of the patient.

[00183] The neoantigen presentation may be modeled as a Poisson Binomial random variable, the utility score can be a probability that a number of presented neoantigens in the treatment subset for the patient is above a minimum threshold.

### **III. Identification of Tumor Specific Mutations in Neoantigens**

[00184] 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.

[00185] 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.

[00186] 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.

**[00187]** 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.

**[00188]** 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.

**[00189]** 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.

**[00190]** 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.

**[00191]** 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.

**[00192]** 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)).

**[00193]** 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.

**[00194]** 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.

**[00195]** 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.

**[00196]** 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.

**[00197]** 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.

**[00198]** 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.

**[00199]** 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.

**[00200]** 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**

**[00201]** 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.

**[00202]** 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.

**[00203]** One or more polypeptides encoded by a neoantigen nucleotide sequence can comprise at least one of: a binding affinity with MHC with an IC<sub>50</sub> 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.

**[00204]** One or more neoantigens can be presented on the surface of a tumor.

**[00205]** One or more neoantigens can be immunogenic in a subject having a tumor, e.g., capable of eliciting a T cell response or a B cell response in the subject.

**[00206]** 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.

**[00207]** 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.

**[00208]** 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.

**[00209]** 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.

**[00210]** 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 IC<sub>50</sub> 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.

**[00211]** In some aspects, neoantigenic peptides and polypeptides do not induce an autoimmune response and/or invoke immunological tolerance when administered to a subject.

**[00212]** 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.

**[00213]** 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).

**[00214]** 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 trichloroacetic 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.

**[00215]** 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.

**[00216]** 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.

**[00217]** 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.

**[00218]** 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**

**[00219]** 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.

**[00220]** 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.

**[00221]** 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.

**[00222]** The vaccine composition can be capable of raising a specific cytotoxic T-cells response and/or a specific helper T-cell response.

**[00223]** 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.

**[00224]** 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.

**[00225]** 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.

**[00226]** 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).

**[00227]** 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.

**[00228]** 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, sorafenib, 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).

**[00229]** 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.

**[00230]** 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 diphtheria toxoid are suitable carriers. Alternatively, the carrier can be dextrans for example sepharose.

**[00231]** 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.

**[00232]** 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**

[00233] Truncal peptides, meaning those presented by all or most tumor subclones, will be prioritized for inclusion into the vaccine.<sup>53</sup> 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.<sup>54</sup>

##### **IV.A.2. Neoantigen prioritization**

[00234] After all of the above 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**

**[00235]** 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.

**[00236]** 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.

**[00237]** A neoantigen can be administered in an amount sufficient to induce a CTL response.

**[00238]** 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.

**[00239]** 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.

**[00240]** 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.

**[00241]** 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.

**[00242]** 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.

**[00243]** 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.

**[00244]** 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.

**[00245]** 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.

**[00246]** 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.

**[00247]** 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.

**[00248]** 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.

**[00249]** 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).

**[00250]** 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.

**[00251]** 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 be cloned into a desired expression vector.

[00252] 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.

[00253] 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.

[00254] 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.

[00255] 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.**

[00256] Research methods for NGS analysis of tumor and normal exome and transcriptomes have been described and applied in the neoantigen identification space.<sup>6,14,15</sup> 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**

[00257] 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<sup>16</sup> 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<sup>17</sup>
  - 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<sup>18</sup>
  - 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<sup>19</sup>, with the same or similar probes used to capture exomes in DNA.

### **VI.A.2. NGS data analysis optimizations**

[00258] 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<sup>20</sup> by merging results from different programs<sup>5</sup>
  - 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<sup>21</sup> and Mutect<sup>22</sup>; and programs that incorporate tumor DNA, tumor RNA and normal DNA, such as UNCEqR, which is particularly advantageous in low-purity samples<sup>23</sup>.
  - b. Indels will be determined with programs that perform local re-assembly, such as Strelka and ABRA<sup>24</sup>.
  - c. Structural rearrangements will be determined using dedicated tools such as Pindel<sup>25</sup> or Breakseq<sup>26</sup>.
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<sup>27</sup>.
  - c. Removal of variants stemming from recurrent sequencing artifacts, even if not observed in the corresponding normal<sup>27</sup>. Examples include variants primarily detected on one strand.
  - d. Removal of variants detected in an unrelated set of controls<sup>27</sup>
5. Accurate HLA calling from normal exome using one of seq2HLA<sup>28</sup>, ATHLATES<sup>29</sup> or Optitype and also combining exome and RNA sequencing data<sup>28</sup>. Additional potential optimizations include the adoption of a dedicated assay for HLA typing such as long-read DNA sequencing<sup>30</sup>, or the adaptation of a method for joining RNA fragments to retain continuity<sup>31</sup>.
6. Robust detection of neo-ORFs arising from tumor-specific splice variants will be performed by assembling transcripts from RNA-seq data using CLASS<sup>32</sup>,

Bayesembler<sup>33</sup>, StringTie<sup>34</sup> 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<sup>35</sup> 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<sup>36</sup> and MAMBA<sup>37</sup>, with mutant sequences re-introduced. Gene expression will be determined with a tool such as Cufflinks<sup>35</sup> 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<sup>38</sup> or HTSeq<sup>39</sup>. 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<sup>40</sup>, the second a uveal melanoma cell line<sup>41</sup>, and the third breast cancer patients<sup>42</sup>.
  - 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<sup>43</sup> (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)

**[00259]** In samples with poly-adenylated RNA, the presence of viral and microbial RNA in the RNA-seq data will be assessed using RNA CoMPASS<sup>44</sup> or a similar method, toward the identification of additional factors that may predict patient response.

#### **VI.B. Isolation and Detection of HLA Peptides**

**[00260]** Isolation of HLA-peptide molecules was performed using classic immunoprecipitation (IP) methods after lysis and solubilization of the tissue sample<sup>55-58</sup>. A clarified lysate was used for HLA specific IP.

**[00261]** 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.<sup>59, 60</sup> 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 |

**[00262]** 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.

[00263] 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.

[00264] MS2 spectra from each analysis are searched against a protein database using Comet<sup>61, 62</sup> and the peptide identification are scored using Percolator<sup>63-65</sup>. 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<sup>75</sup>.

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

[00265] 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 Figure 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**

[00266] 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.

[00267] The presentation identification system 160 is one or computer models, embodied in a computing system as discussed below with respect to FIG. 14, 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.

[00268] 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. 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.

## **VII.B. Presentation Information**

[00269] FIG. 2 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**

[00270] 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.

[00271] 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-DRB1:11:01 and 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.

**[00272]** 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.

**[00273]** 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  $IC_{50} > 1000nm$  are presented by the MHC, and lower  $IC_{50}$  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.

**[00274]** 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.

**[00275]** 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.

**[00276]** 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.

**[00277]** 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.

**[00278]** 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).

**[00279]** 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.

**[00280]** 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.

**[00281]** 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.

**[00282]** 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.

**[00283]** Allele-interacting information can also include the protein sequence of the particular MHC allele.

**[00284]** 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**

**[00285]** 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.

**[00286]** 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 in reference to FIG. 13H, 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).

**[00287]** Allele-noninteracting information can also include the N-terminal sequences flanking the peptide within its source protein sequence.

**[00288]** 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.

**[00289]** Allele-noninteracting information can also include the tissue type, cell type or tumor type of cells of origin of the peptide sequence.

**[00290]** 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.

**[00291]** 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.

**[00292]** 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.

**[00293]** 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.

**[00294]** 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.

**[00295]** 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.

**[00296]** 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.

**[00297]** 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.

**[00298]** 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.

**[00299]** Allele-noninteracting information can also include features describing the presence or absence of a presentation hotspot at the position of the peptide in the source protein of the peptide.

**[00300]** 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).

**[00301]** 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.

**[00302]** 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).

**[00303]** 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.

**[00304]** 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.

**[00305]** 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.

**[00306]** 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.

**[00307]** 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)

[00308] Allele-noninteracting information can also include tumor type (e.g., NSCLC, melanoma).

[00309] 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>.

[00310] Allele-noninteracting information can also include clinical tumor subtype (e.g., squamous lung cancer vs. non-squamous).

[00311] Allele-noninteracting information can also include smoking history.

[00312] Allele-noninteracting information can also include history of sunburn, sun exposure, or exposure to other mutagens.

[00313] 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.

[00314] 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.

[00315] 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**

[00316] 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**

**[00317]** 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  $\mathbf{z}^i$  that include at least a presented or non-presented peptide sequence  $\mathbf{p}^i$ , one or more associated MHC alleles  $\mathbf{a}^i$  associated with the peptide sequence  $\mathbf{p}^i$ , 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.

**[00318]** In one particular implementation referred throughout the remainder of the specification, the dependent variable  $y^i$  is a binary label indicating whether peptide  $\mathbf{p}^i$  was presented by the one or more associated MHC alleles  $\mathbf{a}^i$ . 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  $\mathbf{z}^i$ . 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.

**[00319]** The peptide sequence  $\mathbf{p}^i$  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  $\mathbf{p}^i$  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  $\mathbf{a}^i$  for data instance  $i$  indicate which MHC alleles were present in association with the corresponding peptide sequence  $\mathbf{p}^i$ .

**[00320]** The data management module 312 may also include additional allele-interacting variables, such as binding affinity  $\mathbf{b}^i$  and stability  $\mathbf{s}^i$  predictions in conjunction with the peptide sequences  $\mathbf{p}^i$  and associated MHC alleles  $\mathbf{a}^i$  contained in the training data 170. For example, the training data 170 may contain binding affinity predictions  $\mathbf{b}^i$  between a peptide  $\mathbf{p}^i$  and each of the associated MHC molecules indicated in  $\mathbf{a}^i$ . As another example, the training data 170 may contain stability predictions  $\mathbf{s}^i$  for each of the MHC alleles indicated in  $\mathbf{a}^i$ .

**[00321]** 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  $p^i$ .

**[00322]** 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.

**[00323]** 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.

**[00324]** 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, FIEUHFWDI, and FEWRHRJTRUJR. 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\*01:01 and a peptide sequence QIEJOEIJ. 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.

### **VII.C.2. Encoding Module**

[illegible][illegible]

0 0 0 1 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 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]. The C-terminal flanking sequence  $c^i$  or other sequence data can be similarly encoded as described above. Thus, each independent variable or column in the peptide sequence  $p^i$  or  $c^i$  represents presence of a particular amino acid at a particular position of the sequence.

**[00327]** Although the above method of encoding sequence data was described in reference to sequences having amino acid sequences, the method can similarly be extended to other types of sequence data, such as DNA or RNA sequence data, and the like.

**[00328]** The encoding module 314 also encodes the one or more MHC alleles  $a^i$  for data instance  $i$  as a row vector of  $m$  elements, in which each element  $h=1, 2, \dots, m$  corresponds to a unique identified MHC allele. The elements corresponding to the MHC alleles identified for the data instance  $i$  have a value of 1. Otherwise, the remaining elements have a value of 0. As an example, the alleles HLA-B\*07:02 and HLA-DRB1\*10:01 for a data instance  $i$  corresponding to a multiple-allele cell line among  $m=4$  unique identified MHC allele types {HLA-A\*01:01, HLA-C\*01:08, HLA-B\*07:02, HLA-DRB1\*10:01} may be represented by the row vector of 4 elements  $a^i=[0 \ 0 \ 1 \ 1]$ , in which  $a_3^i=1$  and  $a_4^i=1$ . Although the example is described herein with 4 identified MHC allele types, the number of MHC allele types can be hundreds or thousands in practice. As previously discussed, each data instance  $i$  typically contains at most 6 different MHC allele types in association with the peptide sequence  $p_i$ .

**[00329]** The encoding module 314 also encodes the label  $y_i$  for each data instance  $i$  as a binary variable having values from the set of {0, 1}, in which a value of 1 indicates that peptide  $x^i$  was presented by one of the associated MHC alleles  $a^i$ , and a value of 0 indicates that peptide  $x^i$  was not presented by any of the associated MHC alleles  $a^i$ . When the dependent variable  $y_i$  represents the mass spectrometry ion current, the encoding module 314 may additionally scale the values using various functions, such as the log function having a range of  $(-\infty, \infty)$  for ion current values between  $[0, \infty)$ .

**[00330]** 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).

**[00331]** 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  $\mathbf{x}_h^i$ .

**[00332]** 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  $\mathbf{x}_h^i$ ,

**[00333]** 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  $\mathbf{x}_h^i$ .

**[00334]** In one instance, for peptides presented by class I MHC molecules, the encoding module 314 represents peptide length as a vector  $\mathbf{T}_k = [\mathbb{1}(L_k=8) \mathbb{1}(L_k=9) \mathbb{1}(L_k=10) \mathbb{1}(L_k=11) \mathbb{1}(L_k=12) \mathbb{1}(L_k=13) \mathbb{1}(L_k=14) \mathbb{1}(L_k=15)]$  where  $\mathbb{1}$  is the indicator function, and  $L_k$  denotes the length of peptide  $\mathbf{p}_k$ . The vector  $\mathbf{T}_k$  can be included in the allele-interacting variables  $\mathbf{x}_h^i$ . In another instance, for peptides presented by class II MHC molecules, the encoding module 314 represents peptide length as a vector  $\mathbf{T}_k = [\mathbb{1}(L_k=6) \mathbb{1}(L_k=7) \mathbb{1}(L_k=8) \mathbb{1}(L_k=9) \mathbb{1}(L_k=10) \mathbb{1}(L_k=11) \mathbb{1}(L_k=12) \mathbb{1}(L_k=13) \mathbb{1}(L_k=14) \mathbb{1}(L_k=15) \mathbb{1}(L_k=16) \mathbb{1}(L_k=17) \mathbb{1}(L_k=18) \mathbb{1}(L_k=19) \mathbb{1}(L_k=20) \mathbb{1}(L_k=21) \mathbb{1}(L_k=22) \mathbb{1}(L_k=23) \mathbb{1}(L_k=24) \mathbb{1}(L_k=25) \mathbb{1}(L_k=26) \mathbb{1}(L_k=27) \mathbb{1}(L_k=28) \mathbb{1}(L_k=29) \mathbb{1}(L_k=30)]$  where  $\mathbb{1}$  is the indicator function, and  $L_k$  denotes the length of peptide  $\mathbf{p}_k$ . The vector  $\mathbf{T}_k$  can be included in the allele-interacting variables  $\mathbf{x}_h^i$ .

**[00335]** 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  $\mathbf{x}_h^i$ .

**[00336]** Similarly, the encoding module 314 may represent the allele-noninteracting variables  $\mathbf{w}^i$  as a row vector in which numerical representations of allele-noninteracting variables are concatenated one after the other. For example,  $\mathbf{w}^i$  may be a row vector equal to  $[\mathbf{c}^i]$  or  $[\mathbf{c}^i \mathbf{m}^i \mathbf{w}^i]$  in which  $\mathbf{w}^i$  is a row vector representing any other allele-noninteracting variables in addition to the C-terminal flanking sequence of peptide  $\mathbf{p}^i$  and the mRNA quantification measurement  $\mathbf{m}^i$  associated with the peptide. Alternatively, one or more combination of allele-noninteracting variables may be stored individually (e.g., as individual vectors or matrices).

**[00337]** 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  $w^i$ .

**[00338]** In one instance, the encoding module 314 represents length of source protein or isoform by incorporating the protein length in the allele-noninteracting variables  $w^i$ .

**[00339]** 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  $w^i$ .

**[00340]** 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  $w^i$ .

**[00341]** 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  $w^i$ .

**[00342]** 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  $w^i$ .

**[00343]** 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  $w^i$ .

**[00344]** 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  $w^i$ .

**[00345]** 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  $w^i$ .

**[00346]** 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  $w^i$ .

**[00347]** 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  $w^i$ .

**[00348]** 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  $w^i$ .

**[00349]** 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.

**[00350]** 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 onehot-encoded variables can be included in the allele-noninteracting variables  $w^i$ .

**[00351]** 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  $w^i$ . Alternatively, smoking history can be encoded as a length-one one-hot-enocded 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.

**[00352]** 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  $w^i$ . 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.

**[00353]** 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, we can include 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  $w^i$ , 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.

**[00354]** 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  $w^i$ .

**[00355]** 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  $w^i$ . In another instance, the encoding module 314 represents residue-level annotations of the source protein for peptide  $p^i$  by including an indicator variable, that is equal to 1 if peptide  $p^i$  overlaps with a helix motif and 0 otherwise, or that is equal to 1 if peptide  $p^i$  is completely contained within a helix motif in the allele-noninteracting variables  $w^i$ . In another instance, a feature representing proportion of residues in peptide  $p^i$  that are contained within a helix motif annotation can be included in the allele-noninteracting variables  $w^i$ .

**[00356]** 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  $o_i^k$  is 1 if peptide  $p^k$  comes from protein  $i$  and 0 otherwise.

**[00357]** In one instance, the encoding module 314 represents the source gene  $G=\text{gene}(p^i)$  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, ...,  $L$ .

**[00358]** In one instance, the encoding module 314 represents the tissue type, cell type, tumor type, or tumor histology type  $T=\text{tissue}(p^i)$  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.

[00359] The encoding module 314 may also represent the overall set of variables  $\mathbf{z}^i$  for peptide  $\mathbf{p}^i$  and an associated MHC allele  $h$  as a row vector in which numerical representations of the allele-interacting variables  $\mathbf{x}^i$  and the allele-noninteracting variables  $\mathbf{w}^i$  are concatenated one after the other. For example, the encoding module 314 may represent  $\mathbf{z}_h^i$  as a row vector equal to  $[\mathbf{x}_h^i \ \mathbf{w}^i]$  or  $[\mathbf{w}^i \ \mathbf{x}_h^i]$ .

### **VIII. Training Module**

[00360] 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  $\mathbf{p}^k$  and a set of MHC alleles  $\mathbf{a}^k$  associated with the peptide sequence  $\mathbf{p}^k$ , each presentation model generates an estimate  $u_k$  indicating a likelihood that the peptide sequence  $\mathbf{p}^k$  will be presented by one or more of the associated MHC alleles  $\mathbf{a}^k$ .

#### **VIII.A. Overview**

[00361] 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  $(y_{i \in S}, u_{i \in S}, \theta)$  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:

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

[00362] The presentation model may be a parametric model in which one or more parameters  $\theta$  mathematically specify the dependence between the independent variables and dependent variables. Typically, various parameters of parametric-type presentation models that minimize the loss function  $(y_{i \in S}, u_{i \in S}, \theta)$  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**

[00363] 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.

[00364] In one implementation, the training module 316 models the estimated presentation likelihood  $u_k$  for peptide  $p^k$  for a specific allele  $h$  by:

$$u_k^h = \Pr(p^k \text{ presented; MHC allele } h) = f(g_h(x_h^k; \theta_h)), \quad (2)$$

where peptide sequence  $x_h^k$  denotes the encoded allele-interacting variables for peptide  $p^k$  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  $x_h^k$  based on a set of parameters  $\theta_h$  determined for MHC allele  $h$ . The values for the set of parameters  $\theta_h$  for each MHC allele  $h$  can be determined by minimizing the loss function with respect to  $\theta_h$ , where  $i$  is each instance in the subset  $S$  of training data 170 generated from cells expressing the single MHC allele  $h$ .

[00365] 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.

**[00366]** 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:

$$f(z) = \frac{\exp(z)}{1 + \exp(z)}. \quad (4)$$

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

$$f(z) = \tanh(z) \quad (5)$$

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.

**[00367]** 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**

**[00368]** In one particular implementation referred throughout the specification, the dependency function  $g_h(\cdot)$  is an affine function given by:

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

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

**[00369]** In another particular implementation referred throughout the specification, the dependency function  $g_h(\cdot)$  is a network function given by:

$$g_h(x_h^i; \theta_h) = NN_h(x_h^i; \theta_h). \quad (7)$$

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  $\theta_h$ . 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.

**[00370]** 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.

**[00371]** In one instance referred throughout the remainder of the specification, each MHC allele in  $h=1, 2, \dots, m$  is associated with a separate network model, and  $NN_h(\cdot)$  denotes the output(s) from a network model associated with MHC allele  $h$ .

**[00372]** 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  $\theta_3(1), \theta_3(2), \dots, \theta_3(10)$ . 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.

**[00373]** In another instance, the identified MHC alleles  $h=1, 2, \dots, m$  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  $\theta_h$  may correspond to a set of parameters for the single network model, and thus, the set of parameters  $\theta_h$  may be shared by all MHC alleles.

**[00374]** FIG. 6A illustrates an example network model  $NN_H(\cdot)$  shared by MHC alleles  $h=1, 2, \dots, m$ . As shown in FIG. 6A, 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  $h=3$ .

[00375] In yet another instance, the single network model  $NN_H(\cdot)$  may be a network model that outputs a dependency score given the allele interacting variables  $\mathbf{x}_h^k$  and the encoded protein sequence  $\mathbf{d}_h$  of an MHC allele  $h$ . In such an instance, the set of parameters  $\theta_h$  may again correspond to a set of parameters for the single network model, and thus, the set of parameters  $\theta_h$  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 inputs  $[\mathbf{x}_h^k \mathbf{d}_h]$  to the single network model. 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 their protein sequence.

[00376] FIG. 6B illustrates an example network model  $NN_H(\cdot)$  shared by MHC alleles. As shown in FIG. 6B, the network model  $NN_H(\cdot)$  receives the allele interacting variables and protein sequence of MHC allele  $h=3$  as input, and outputs a dependency score  $NN_3(\mathbf{x}_3^k)$  corresponding to the MHC allele  $h=3$ .

[00377] In yet another instance, the dependency function  $g_h(\cdot)$  can be expressed as:

$$g_h(\mathbf{x}_h^k; \theta_h) = g'_h(\mathbf{x}_h^k; \theta'_h) + \theta_h^0$$

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

[00378] In another implementation, the bias parameter  $\theta_h^0$  may be shared according to the gene family of the MHC allele  $h$ . That is, the bias parameter  $\theta_h^0$  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  $\theta_h^0$  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  $\theta_h^0$  for each of these MHC alleles may be shared.

[00379] Returning to equation (2), as an example, the likelihood that peptide  $\mathbf{p}^k$  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:

$$u_k^3 = f(\mathbf{x}_3^k \cdot \theta_3),$$

where  $\mathbf{x}_3^k$  are the identified allele-interacting variables for MHC allele  $h=3$ , and  $\theta_3$  are the set of parameters determined for MHC allele  $h=3$  through loss function minimization.

[00380] 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  $\theta_3$  are the set of parameters determined for the network model  $NN_3(\cdot)$  associated with MHC allele  $h=3$ .

[00381] 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**

[00382] In one implementation, the training module 316 incorporates allele-noninteracting variables and models the estimated presentation likelihood  $u_k$  for peptide  $p^k$  by:

$$u_k^h = \Pr(p^k \text{ presented}) = f(g_w(w^k; \theta_w) + g_h(x_h^i; \theta_h)), \quad (8)$$

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  $\theta_w$  determined for the allele-noninteracting variables. Specifically, the values for the set of parameters  $\theta_h$  for each MHC allele  $h$  and the set of parameters  $\theta_w$  for allele-noninteracting variables can be determined by minimizing the loss function with respect to  $\theta_h$  and  $\theta_w$ , where  $i$  is each instance in the subset  $S$  of training data 170 generated from cells expressing single MHC alleles.

[00383] 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  $p^k$ .

[00384] According to equation (8), 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  $h$ .

[00385] 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:

$$u_k^h = \Pr(p^k \text{ presented; allele } h) = f(g_h([x_h^k w^k]; \theta_h)). \quad (9)$$

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

[00386] 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  $w^k$ .

[00387] Specifically, the dependency function  $g_w(\cdot)$  is an affine function given by:

$$g_w(w^k; \theta_w) = w^k \cdot \theta_w.$$

that linearly combines the allele-noninteracting variables in  $w^k$  with a corresponding parameter in the set of parameters  $\theta_w$ .

[00388] The dependency function  $g_w(\cdot)$  may also be a network function given by:

$$g_h(w^k; \theta_w) = NN_w(w^k; \theta_w).$$

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

[00389] In another instance, the dependency function  $g_w(\cdot)$  for the allele-noninteracting variables can be given by:

$$g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + h(m^k; \theta_w^m), \quad (10)$$

where  $g'_w(w^k; \theta'_w)$  is the affine function, the network function with the set of allele noninteracting parameters  $\theta'_w$ , or the like,  $m^k$  is the mRNA quantification measurement for peptide  $p^k$ ,  $h(\cdot)$  is a function transforming the quantification measurement, and  $\theta_w^m$  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.

**[00390]** In yet another instance, the dependency function  $g_w(\cdot)$  for the allele-noninteracting variables can be given by:

$$g_w(\mathbf{w}^k; \boldsymbol{\theta}_w) = g'_w(\mathbf{w}^k; \boldsymbol{\theta}'_w) + \boldsymbol{\theta}_w^o \cdot \mathbf{o}^k, \quad (11)$$

where  $g'_w(\mathbf{w}^k; \boldsymbol{\theta}'_w)$  is the affine function, the network function with the set of allele noninteracting parameters  $\boldsymbol{\theta}'_w$ , or the like,  $\mathbf{o}^k$  is the indicator vector described in Section VII.C.2 representing proteins and isoforms in the human proteome for peptide  $\mathbf{p}^k$ , 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  $\mathbf{o}^k$  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  $\lambda$  can be determined through appropriate methods.

**[00391]** In yet another instance, the dependency function  $g_w(\cdot)$  for the allele-noninteracting variables can be given by:

$$g_w(\mathbf{w}^k; \boldsymbol{\theta}_w) = g'_w(\mathbf{w}^k; \boldsymbol{\theta}'_w) + \sum_{l=1}^L \mathbb{1}(\text{gene}(\mathbf{p}^k = l)) \cdot \boldsymbol{\theta}_w^l, \quad (12)$$

where  $g'_w(\mathbf{w}^k; \boldsymbol{\theta}'_w)$  is the affine function, the network function with the set of allele noninteracting parameters  $\boldsymbol{\theta}'_w$ , or the like,  $\mathbb{1}(\text{gene}(\mathbf{p}^k = l))$  is the indicator function that equals to 1 if peptide  $\mathbf{p}^k$  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  $\boldsymbol{\theta}_w^{l=1, 2, \dots, L}$  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  $\lambda$  can be determined through appropriate methods.

**[00392]** In yet another instance, the dependency function  $g_w(\cdot)$  for the allele-noninteracting variables can be given by:

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

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

**[00393]** In one variation, when  $L$  or  $M$  is significantly high, and thus, the number of parameters  $\boldsymbol{\theta}_w^{lm=l, 2, \dots, LM}$  are significantly high, a parameter regularization term, such as  $\lambda \cdot \|\boldsymbol{\theta}_w^{lm}\|$ , 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  $\lambda$  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 coefficients for the same source gene do not significantly differ between tissue types. For example, a penalization term such as:

$$\lambda \cdot \sum_{l=1}^L \sqrt{\sum_{m=1}^M (\boldsymbol{\theta}_w^{lm} - \overline{\boldsymbol{\theta}_w^l})^2}$$

where  $\overline{\boldsymbol{\theta}_w^l}$  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.

**[00394]** In practice, the additional terms of any of equations (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 (10) and the term indicating source gene antigenicity in equation (12) may be summed together along with any other affine or network function to generate the dependency function for allele noninteracting variables.

**[00395]** Returning to equation (8), as an example, the likelihood that peptide  $\mathbf{p}^k$  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:

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

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

[00396] As another example, the likelihood that peptide  $\mathbf{p}^k$  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:

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

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

[00397] FIG. 8 illustrates generating a presentation likelihood for peptide  $\mathbf{p}^k$  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  $\mathbf{x}_3^k$  for MHC allele  $h=3$  and generates the output  $NN_3(\mathbf{x}_3^k)$ . The network model  $NN_w(\cdot)$  receives the allele-noninteracting variables  $\mathbf{w}^k$  for peptide  $\mathbf{p}^k$  and generates the output  $NN_w(\mathbf{w}^k)$ . The outputs are combined and mapped by function  $f(\cdot)$  to generate the estimated presentation likelihood  $u_k$ .

### **VIII.C. Multiple-Allele Models**

[00398] 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**

[00399] In one implementation, the training module 316 models the estimated presentation likelihood  $u_k$  for peptide  $\mathbf{p}^k$  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)-(11). Specifically, the presentation likelihood  $u_k$  can be any function of  $u_k^{h \in H}$ . In one implementation, as shown in equation (12), 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  $H$ .

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

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

**[00400]** In one implementation, the training module 316 models the estimated presentation likelihood  $u_k$  for peptide  $p^k$  by:

$$u_k = \Pr(p^k \text{ presented}) = f\left(\sum_{h=1}^m a_h^k \cdot g_h(x_h^k; \theta_h)\right), \quad (13)$$

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  $\theta_h$  for each MHC allele  $h$  can be determined by minimizing the loss function with respect to  $\theta_h$ , 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.

**[00401]** 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  $H$ .

**[00402]** 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  $p^k$ .

**[00403]** 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:

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

where  $x_2^k, x_3^k$  are the identified allele-interacting variables for MHC alleles  $h=2, h=3$ , and  $\theta_2, \theta_3$  are the set of parameters determined for MHC alleles  $h=2, h=3$ .

**[00404]** 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:

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

where  $NN_2(\cdot)$ ,  $NN_3(\cdot)$  are the identified network models for MHC alleles  $h=2$ ,  $h=3$ , and  $\theta_2$ ,  $\theta_3$  are the set of parameters determined for MHC alleles  $h=2$ ,  $h=3$ .

[00405] 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**

[00406] In one implementation, the training module 316 incorporates allele-noninteracting variables and models the estimated presentation likelihood  $u_k$  for peptide  $p^k$  by:

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

where  $w^k$  denotes the encoded allele-noninteracting variables for peptide  $p^k$ . Specifically, the values for the set of parameters  $\theta_h$  for each MHC allele  $h$  and the set of parameters  $\theta_w$  for allele-noninteracting variables can be determined by minimizing the loss function with respect to  $\theta_h$  and  $\theta_w$ , 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.

[00407] 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  $H$ .

[00408] 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  $p^k$ .

[00409] 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:

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

where  $\mathbf{w}^k$  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.

[00410] 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:

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

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

[00411] 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  $\mathbf{x}_2^k$  for MHC allele  $h=2$  and generates the output  $NN_2(\mathbf{x}_2^k)$ . The network model  $NN_3(\cdot)$  receives the allele-interacting variables  $\mathbf{x}_3^k$  for MHC allele  $h=3$  and generates the output  $NN_3(\mathbf{x}_3^k)$ . The network model  $NN_w(\cdot)$  receives the allele-noninteracting variables  $\mathbf{w}^k$  for peptide  $p^k$  and generates the output  $NN_w(\mathbf{w}^k)$ . The outputs are combined and mapped by function  $f(\cdot)$  to generate the estimated presentation likelihood  $u_k$ .

[00412] Alternatively, the training module 316 may include allele-noninteracting variables  $\mathbf{w}^k$  in the prediction by adding the allele-noninteracting variables  $\mathbf{w}^k$  to the allele-interacting variables  $\mathbf{x}_h^k$  in equation (15). Thus, the presentation likelihood can be given by:

$$u_k = \Pr(p^k \text{ presented}) = f\left(\sum_{h=1}^m a_h^k \cdot g_h([\mathbf{x}_h^k \mathbf{w}^k]; \boldsymbol{\theta}_h)\right). \quad (15)$$

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

[00413] In another implementation, the training module 316 models the estimated presentation likelihood  $u_k$  for peptide  $p^k$  by:

$$u_k = \Pr(\mathbf{p}^k \text{ presented}) = r \left( s(\mathbf{v} = [a_1^k \cdot u_k'^1(\boldsymbol{\theta}) \dots a_m^k \cdot u_k'^m(\boldsymbol{\theta})]) \right), \quad (16)$$

where elements  $a_h^k$  are 1 for the multiple MHC alleles  $h \in H$  associated with peptide sequence  $\mathbf{p}^k$ ,  $u_k'^h$  is an implicit per-allele presentation likelihood for MHC allele  $h$ , vector  $\mathbf{v}$  is a vector in which element  $v_h$  corresponds to  $a_h^k \cdot u_k'^h$ ,  $s(\cdot)$  is a function mapping the elements of  $\mathbf{v}$ , 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  $\boldsymbol{\theta}$  for the implicit per-allele likelihoods can be determined by minimizing the loss function with respect to  $\boldsymbol{\theta}$ , 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.

**[00414]** The presentation likelihood in the presentation model of equation (17) is modeled as a function of implicit per-allele presentation likelihoods  $u_k'^h$  that each correspond to the likelihood peptide  $\mathbf{p}^k$  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  $\mathbf{p}^k$  will be presented by a set of MHC alleles  $H$  as a whole, but can also provide individual likelihoods  $u_k'^{h \in H}$  that indicate which MHC allele  $h$  most likely presented peptide  $\mathbf{p}^k$ . An advantage of this is that the presentation model can generate the implicit likelihoods without training data for cells expressing single MHC alleles.

**[00415]** 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

[00416] In one particular implementation,  $s(\cdot)$  is a summation function, and the presentation likelihood is given by summing the implicit per-allele presentation likelihoods:

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

[00417] In one implementation, the implicit per-allele presentation likelihood for MHC allele  $h$  is generated by:

$$u_k'^h = f(g_h(\mathbf{x}_h^k; \boldsymbol{\theta}_h)), \quad (18)$$

such that the presentation likelihood is estimated by:

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

[00418] According to equation (19), the presentation likelihood that a peptide sequence  $\mathbf{p}^k$  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  $\mathbf{p}^k$  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  $u_k'^h$ . The per-allele likelihoods  $u_k'^h$  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  $\mathbf{p}^k$  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.

[00419] As an example, the likelihood that peptide  $\mathbf{p}^k$  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:

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

where  $\mathbf{x}_2^k, \mathbf{x}_3^k$  are the identified allele-interacting variables for MHC alleles  $h=2, h=3$ , and  $\boldsymbol{\theta}_2, \boldsymbol{\theta}_3$  are the set of parameters determined for MHC alleles  $h=2, h=3$ .

[00420] As another example, the likelihood that peptide  $\mathbf{p}^k$  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:

$$u_k = r \left( f(NN_2(\mathbf{x}_2^k; \boldsymbol{\theta}_2)) + f(NN_3(\mathbf{x}_3^k; \boldsymbol{\theta}_3)) \right),$$

where  $NN_2(\cdot)$ ,  $NN_3(\cdot)$  are the identified network models for MHC alleles  $h=2$ ,  $h=3$ , and  $\theta_2$ ,  $\theta_3$  are the set of parameters determined for MHC alleles  $h=2$ ,  $h=3$ .

[00421] 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. 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  $NN_3(x_3^k)$ . Each output is mapped by function  $f(\cdot)$  and combined to generate the estimated presentation likelihood  $u_k$ .

[00422] 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**

[00423] In one implementation, the implicit per-allele presentation likelihood for MHC allele  $h$  is generated by:

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

such that the presentation likelihood is generated by:

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

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

[00424] 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.

**[00425]** 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:

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

where  $\mathbf{w}^k$  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.

**[00426]** 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:

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

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

**[00427]** 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  $\mathbf{x}_2^k$  for MHC allele  $h=2$  and generates the output  $NN_2(\mathbf{x}_2^k)$ . The network model  $NN_w(\cdot)$  receives the allele-noninteracting variables  $\mathbf{w}^k$  for peptide  $p^k$  and generates the output  $NN_w(\mathbf{w}^k)$ . The outputs are combined and mapped by function  $f(\cdot)$ . The network model  $NN_3(\cdot)$  receives the allele-interacting variables  $\mathbf{x}_3^k$  for MHC allele  $h=3$  and generates the output  $NN_3(\mathbf{x}_3^k)$ , which is again combined with the output  $NN_w(\mathbf{w}^k)$  of the same network model  $NN_w(\cdot)$  and mapped by function  $f(\cdot)$ . Both outputs are combined to generate the estimated presentation likelihood  $u_k$ .

**[00428]** In another implementation, the implicit per-allele presentation likelihood for MHC allele  $h$  is generated by:

$$u_k'^h = f \left( g_h([\mathbf{x}_h^k \mathbf{w}^k]; \boldsymbol{\theta}_h) \right). \quad (22)$$

such that the presentation likelihood is generated by:

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

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

[00429] In one implementation,  $s(\cdot)$  is a second-order function, and the estimated presentation likelihood  $u_k$  for peptide  $p^k$  is given by:

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

where elements  $u_k'^h$  are the implicit per-allele presentation likelihood for MHC allele  $h$ . The values for the set of parameters  $\theta$  for the implicit per-allele likelihoods can be determined by minimizing the loss function with respect to  $\theta$ , 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.

[00430] 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.

[00431] 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  $H$ .

[00432] 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  $\theta_2, \theta_3$  are the set of parameters determined for HLA alleles  $h=2, h=3$ .

[00433] 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:

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

where  $NN_2(\cdot), NN_3(\cdot)$  are the identified network models for HLA alleles  $h=2, h=3$ , and  $\theta_2, \theta_3$  are the set of parameters determined for HLA alleles  $h=2, h=3$ .

### **IX. Example 5: Prediction Module**

**[00434]** 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.

**[00435]** 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**

**[00436]** The patient selection module 324 selects a subset of patients for vaccine treatment based on whether the patients satisfy an 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 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, but may result in a higher proportion of vaccine-treated patients that receive effective treatment (e.g., 1 or more tumor-specific neoantigens (TSNA) are delivered). On the other

hand, a lenient inclusion criteria results in a higher number of patients that will be treated with the vaccine, but may result in a lower proportion of vaccine-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 the vaccine and proportion of patients that receive effective treatment as a result of vaccine treatment.

**[00437]** 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.

**[00438]** 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.

**[00439]** 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.

**[00440]** 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  $p^{i1}, p^{i2}, \dots, p^{iv}$  each having the highest presentation likelihoods  $u_{i1}, u_{i2}, \dots, u_{iv}$ , presentation of neoantigen candidate  $p^{ij}$  is given by random variable  $A_{ij}$ , in which:

$$P(A_{ij} = 1) = u_{ij}, \quad P(A_{ij} = 0) = 1 - u_{ij}. \quad (24)$$

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:

$$\text{util}_i(S_i) = \mathbb{E} \left[ \sum_{j=1}^v A_{ij} \right] = \sum_{j=1}^v u_{ij}. \quad (25)$$

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

**[00441]** 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:

$$N_i = \sum_{j=1}^v A_{ij} \sim \text{PBD}(u_{i1}, u_{i2}, \dots, u_{iv}). \quad (26)$$

where  $\text{PBD}(\cdot)$  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:

$$\text{util}_i(S_i) = \mathbb{P}[N_i \geq k] = \sum_{m=k}^v \mathbb{P}[N_i = m]. \quad (27)$$

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

**[00442]** In another embodiment, the number of neoantigens in the treatment subset for a patient is not required to be limited to the vaccine capacity  $v$ , and the patient selection module 324 may select patients using utility scores determined based on any set of candidate neoantigens of the patient. For example, the utility scores may be determined based on all mutations or candidate neoantigens identified for the patient. The utility scores may be generated using, for example, the methods described in conjunction with equations (24) through (27), where  $v$  is now a variable  $v(i)$  dependent on the patient  $i$ , and indicates the total number of mutations or candidate neoantigens identified for the patient.

**[00443]** 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.

**[00444]** 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 10<sup>th</sup> percentile to the 0.1<sup>th</sup> percentile. Optionally, the utility score may count only those neoantigens detected as expressed via RNA-seq.

**[00445]** 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.

#### **XI. Example 7: Neoantigen Burden for Immune Checkpoint Inhibitor Therapy and Other Immunotherapies**

**[00446]** The patient selection module 324 can also use the utility scores as defined in Section X above to select patients for immune checkpoint inhibitor therapy (e.g., PD-1, CTLA4) or any other immunotherapy where neoantigen burden may be relevant for efficacy. Other immunotherapies may include immunostimulatory agents, immune stimulatory molecule agonists (e.g., CD40), oncolytic viruses (e.g., T-VEC), neoantigen or other cancer antigen-containing therapeutic vaccines, neoantigen or other cancer antigen targeted adoptive

cell therapies, tumor microenvironment modulators (e.g., TGFbeta), or any combinations of these with immune checkpoint inhibitors.

**[00447]** For example, in some embodiments, the immunostimulatory agent is an agent that blocks signaling of an inhibitory receptor of an immune cell, or a ligand thereof. In some aspects, the inhibitory receptor or ligand is selected from CTLA-4, PD-1, PD-L1, LAG-3, Tim3, TIGIT, neuritin, BTLA, KIR, and combinations thereof. In some aspects, the agent is selected from an anti-PD-1 antibody (e.g., pembrolizumab or nivolumab), and anti-PD-L1 antibody (e.g., atezolizumab), an anti-CTLA-4 antibody (e.g., ipilimumab), and combinations thereof. In some aspects, the agent is pembrolizumab. In some aspects, the agent is nivolumab. In some aspects, the agent is atezolizumab.

**[00448]** In some embodiments, the therapeutic agent is an agent that inhibits the interaction between PD-1 and PD-L1. In some aspects, the additional therapeutic agent that inhibits the interaction between PD-1 and PD-L1 is selected from an antibody, a peptidomimetic and a small molecule. In some aspects, the additional therapeutic agent that inhibits the interaction between PD-1 and PD-L1 is selected from pembrolizumab, nivolumab, atezolizumab, avelumab, durvalumab, BMS-936559, sulfamonomethoxine 1, and sulfamethizole 2. In some embodiments, the additional therapeutic agent that inhibits the interaction between PD-1 and PD-L1 is any therapeutic known in the art to have such activity, for example as described in Weinmann et al., Chem Med Chem, 2016, 14:1576 (DOI: 10.1002/cmdc.201500566), incorporated by reference in its entirety.

**[00449]** In some embodiments, the immunostimulatory agent is an agonist of a co-stimulatory receptor of an immune cell. In some aspects, the co-stimulatory receptor is selected from OX40, ICOS, CD27, CD28, 4-1BB, or CD40. In some embodiments, the agonist is an antibody.

**[00450]** In some embodiments, the immunostimulatory agent is a cytokine. In some aspects, the cytokine is selected from IL-2, IL-5, IL-7, IL-12, IL-15, IL-21, and combinations thereof.

**[00451]** In some embodiments, the immunostimulatory agent is an oncolytic virus. In some aspects, the oncolytic virus is selected from a herpes simplex virus, a vesicular stomatitis virus, an adenovirus, a Newcastle disease virus, a vaccinia virus, and a maraba virus.

**[00452]** In some embodiments, the immunostimulatory agent is a T cell with a chimeric antigen receptor (CAR-T cell). In some embodiments, the immunostimulatory agent is a bi-

or multi-specific T cell directed antibody. In some embodiments, the immunostimulatory agent is an anti-TGF- $\beta$  antibody. In some embodiments, the immunostimulatory agent is a TGF- $\beta$  trap.

**[00453]** In some embodiments, the therapeutic agent is a vaccine to a tumor antigen. Any suitable antigen may be targeted by the vaccine, provided that it is present in a tumor treated by the methods provided herein. In some aspects, the tumor antigen is a tumor antigen that is overexpressed in comparison its expression levels in normal tissue. In some aspects, the tumor antigen is selected from cancer testis antigen, differentiation antigen, NY-ESO-1, MAGE-A1, MART, and combinations thereof. In some embodiments, the therapeutic agent is a vaccine to one or more neoantigens. The neoantigens in the vaccine may be identified by the methods provided herein.

**[00454]** Specifically, the patient selection module 324 determines a neoantigen burden that indicates a total expected number of presented neoantigens for each patient. The patients that have neoantigen burdens that satisfy an inclusion criteria may be administered with checkpoint inhibitor therapy. For example, patients that have neoantigen burdens above a predetermined threshold may be administered with the therapy. In one embodiment, the neoantigen burden is the utility score given in Section XI, in which  $v$  is the total number of mutations or candidate neoantigens identified for a patient, rather than a subset of candidate neoantigens for the patient.

**[00455]** When a neoantigen burden is higher in a given tumor relative to a median it can indicate that the subject with that tumor is likely to benefit from treatment with a checkpoint inhibitor such as anti-CTLA4, anti-PD1, and/or anti-PDL1. A neoantigen burden can be a better indicator of benefit from checkpoint inhibitor therapy than mutational burden, for example, because neoantigens are generally more likely to be presented on the tumor cell surface and recognized by T cells that are more active against the tumor following checkpoint inhibitor therapy.

**[00456]** In another embodiment, the patient selection module 324 can use utility scores generated from a combination of one or more of the following features: predicted HLA class I neoantigen burden, predicted HLA class II neoantigen burden, and tumor mutation burden. The predicted HLA class I neoantigen burden for a patient is the neoantigen burden for the set of class I HLA alleles of the patient, and indicates the total expected number of presented neoantigens on the class I HLA alleles of the patient. The predicted HLA class II neoantigen burden for a patient is the neoantigen burden for the set of class II HLA alleles of the patient,

and indicates the total expected number of presented neoantigens on the class II HLA alleles of the patient. For example, the utility score could be computed as  $f(\text{class I neoantigen burden, class II neoantigen burden, tumor mutation burden; } b)$ , where  $f(\cdot)$  is a function parameterized by a set  $b$  of machine-learned parameters. The set  $b$  of machine learned parameters may depend on tumor type (e.g.,  $b$  may be different for melanoma and non-small-cell lung cancer).

**[00457]** In another embodiment, the patient selection module 324 can use utility scores that incorporate information about immunogenic tumor antigens other than neoantigens. Examples of immunogenic tumor antigens other than neoantigens include cancer-germline antigens (CGAs, e.g., MAGEA3), differentiation antigens (e.g., tyrosinase), and antigens overexpressed in tumors (e.g., CEA). The expression levels of such antigens can be determined using at least tumor RNA sequencing data, and the expected number of HLA class I or class II epitopes from such genes presented by the patient's tumor's HLA alleles can be determined using the RNA sequencing data for each gene and applying the presentation model to each peptide from the set of tumor antigens. These presentation likelihoods can be incorporated in a utility score computed as  $f(\text{class I neoantigen burden, class II neoantigen burden, tumor mutation burden, class I non-neoantigen tumor antigen burden, class II non-neoantigen tumor antigen burden; } b)$ , where  $f(\cdot)$  is a function parameterized by a set  $b$  of machine-learned parameters. The set  $b$  of machine learned parameters may depend on tumor type (e.g.,  $b$  may be different for melanoma and non-small-cell lung cancer).

**[00458]** When the utility scores are higher, this indicates a tumor that presents more HLA epitopes that are perceived by the patient's immune system as foreign or non-self. Patients with tumors that present more non-self HLA epitopes can be more likely to benefit from checkpoint inhibitors or other immunotherapies because the tumors are more likely to be recognized by T cells that are more active against the tumor following immunotherapy.

**[00459]** The utility scores described in Section X above can also be adapted to select patients for treatment with adoptive cell-therapy (e.g., expanded TIL, CAR-T, or engineered TCR) by using  $f(\text{class I neoantigen burden, class II neoantigen burden, tumor mutation burden, class I non-neoantigen tumor antigen burden, class II non-neoantigen tumor antigen burden})$  where the Class I and Class II neoantigens and non-neoantigens are taken only as those present or predicted present in the adoptive cell-therapy. For example, in the case of

engineering TCR therapy against a single neoantigen epitope,  $f$  can reduce to the presentation likelihood of that single epitope.

## **XII. Example 8: Experimentation Results Showing Example Patient Selection Performance**

[00460] 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](http://www.ebi.ac.uk/pride/archive/projects/PXD0000394)). As described in more detail below in conjunction with FIG. 13A, 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.

[00461] 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](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.

[00462] As laid out in the following examples and with reference to FIGs. 13A-13G, 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.

**[00463]** 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.

#### **XII.A. Example 8A: Frequency Distribution of Tumor Mutation Burden for NSCLC Cancer Patients**

**[00464]** FIG. 13A 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. 13A shows a range of 3-1786 mutations, in which 30% of the patients have fewer than 100 mutations. Although not shown in FIG. 13A, 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.

**[00465]** 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. 13A 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. 13B through 13G illustrate experimental results on patient selection based on presented neoantigens of the patients in the test set.

**XII.B. Example 8B: Proportion of Selected Patients with Neoantigen Presentation based on Tumor Mutation Burden Inclusion Criteria**

[00466] FIG. 13B illustrates the number of presented neoantigens in simulated vaccines for patients selected based on an inclusion criteria of whether the patients satisfy a minimum tumor mutation burden. The proportion of selected patients that have at least a certain number of presented neoantigens in the corresponding test is identified.

[00467] In FIG. 13B, 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 tumor 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.

[00468] As indicated in FIG. 13B, the proportion of selected patients with presented neoantigens increases significantly with higher tumor mutation burden. This indicates that tumor mutation burden as an inclusion criteria can be effective in selecting patients for whom neoantigen vaccines are more likely to induce successful immune responses.

**XII.C. Example 8C: Comparison of Neoantigen Presentation for Vaccines Identified by Presentation Models vs. State-of-the-Art Models**

[00469] FIG. 13C 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.

**[00470]** In FIG. 13C, 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  $\nu$  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  $\nu$  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.

**[00471]** As indicated in FIG. 13C, 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.

#### **XII.D. Example 8D: Effect of HLA Coverage on Neoantigen Presentation for Vaccines Identified Through Presentation Models**

**[00472]** FIG. 13D 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.

[00473] In FIG. 13D, 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).

[00474] As indicated in FIG. 13D, 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.

#### **XII.E. Example 8E: Comparison of Neoantigen Presentation for Patients Selected by Tumor Mutation Burden vs. Expected Number of Presented Neoantigens**

[00475] FIG. 13E compares the number of presented neoantigens in simulated vaccines between patients selected based on tumor 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.

[00476] In FIG. 13E, 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 tumor 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 tumor 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).

[00477] As indicated in FIG. 13E, patients selected based on expectation utility scores receive a vaccine containing presented neoantigens at a higher rate than patients selected based on tumor mutation burden. However, patients selected based on tumor mutation burden receive a vaccine containing presented neoantigens at a higher rate than unselected patients. Thus, tumor mutation burden is an effective patient selection criteria for successful neoantigen vaccine treatment, though expectation utility scores are more effective.

### **XIII. Example Computer**

[00478] FIG. 14 illustrates an example computer 1400 for implementing the entities shown in FIGS. 1 and 3. The computer 1400 includes at least one processor 1402 coupled to a chipset 1404. The chipset 1404 includes a memory controller hub 1420 and an input/output (I/O) controller hub 1422. A memory 1406 and a graphics adapter 1412 are coupled to the memory controller hub 1420, and a display 1418 is coupled to the graphics adapter 1412. A storage device 1408, an input device 1414, and network adapter 1416 are coupled to the I/O controller hub 1422. Other embodiments of the computer 1400 have different architectures.

[00479] The storage device 1408 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 1406 holds instructions and data used by the processor 1402. The input interface 1414 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 1400. In some embodiments, the computer 1400 may be configured to receive input (e.g., commands) from the input interface 1414 via gestures from the user. The graphics adapter 1412 displays images and other information on the display 1418. The network adapter 1416 couples the computer 1400 to one or more computer networks.

**[00480]** The computer 1400 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 1408, loaded into the memory 1406, and executed by the processor 1402.

**[00481]** The types of computers 1400 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 1400 or multiple computers 1400 communicating with each other through a network such as in a server farm. The computers 1400 can lack some of the components described above, such as graphics adapters 1412, and displays 1418.

## References

1. Desrichard, A., Snyder, A. & Chan, T. A. Cancer Neoantigens and Applications for Immunotherapy. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* (2015). doi:10.1158/1078-0432.CCR-14-3175
2. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
3. Gubin, M. M., Artyomov, M. N., Mardis, E. R. & Schreiber, R. D. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* **125**, 3413–3421 (2015).
4. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
5. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
6. Carreno, B. M. *et al.* Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348**, 803–808 (2015).
7. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **344**, 641–645 (2014).
8. Hacohe, N. & Wu, C. J.-Y. United States Patent Application: 0110293637 - COMPOSITIONS AND METHODS OF IDENTIFYING TUMOR SPECIFIC NEOANTIGENS. (A1). at <<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1=20110293637.PG.NR.>>
9. Lundegaard, C., Hoof, I., Lund, O. & Nielsen, M. State of the art and challenges in sequence based T-cell epitope prediction. *Immunome Res.* **6 Suppl 2**, S3 (2010).
10. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576 (2014).
11. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics MCP* **14**, 658–673 (2015).
12. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).
13. Yoshida, K. & Ogawa, S. Splicing factor mutations and cancer. *Wiley Interdiscip. Rev. RNA* **5**, 445–459 (2014).
14. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
15. Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **124**, 453–462 (2014).
16. Downing, S. R. *et al.* United States Patent Application: 0120208706 - OPTIMIZATION OF MULTIGENE ANALYSIS OF TUMOR SAMPLES. (A1). at <<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1=20120208706.PG.NR.>>
17. Target Capture for NextGen Sequencing - IDT. at <<http://www.idtdna.com/pages/products/nextgen/target-capture>>

18. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
19. Cieslik, M. *et al.* The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).
20. Bodini, M. *et al.* The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations. *Blood* **125**, 600–605 (2015).
21. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28**, 1811–1817 (2012).
22. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
23. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, e107 (2014).
24. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinforma. Oxf. Engl.* **30**, 2813–2815 (2014).
25. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinforma. Oxf. Engl.* **25**, 2865–2871 (2009).
26. Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010).
27. Frampton, G. M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
28. Boegel, S. *et al.* HLA typing from RNA-Seq sequence reads. *Genome Med.* **4**, 102 (2012).
29. Liu, C. *et al.* ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* **41**, e142 (2013).
30. Mayor, N. P. *et al.* HLA Typing for the Next Generation. *PloS One* **10**, e0127153 (2015).
31. Roy, C. K., Olson, S., Graveley, B. R., Zamore, P. D. & Moore, M. J. Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation. *eLife* **4**, (2015).
32. Song, L. & Florea, L. CLASS: constrained transcript assembly of RNA-seq reads. *BMC Bioinformatics* **14 Suppl 5**, S14 (2013).
33. Maretty, L., Sibbesen, J. A. & Krogh, A. Bayesian transcriptome assembly. *Genome Biol.* **15**, 501 (2014).
34. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
35. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinforma. Oxf. Engl.* (2011). doi:10.1093/bioinformatics/btr355
36. Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* **15**, 81 (2014).
37. Rivas, M. A. *et al.* Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
38. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21**, 1728–1737 (2011).

39. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinforma. Oxf. Engl.* **31**, 166–169 (2015).
40. Furney, S. J. *et al.* SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov.* (2013). doi:10.1158/2159-8290.CD-13-0330
41. Zhou, Q. *et al.* A chemical genetics approach for the functional assessment of novel cancer genes. *Cancer Res.* (2015). doi:10.1158/0008-5472.CAN-14-2930
42. Maguire, S. L. *et al.* SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J. Pathol.* **235**, 571–580 (2015).
43. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking* **13**, 311–319 (2015).
44. Xu, G. *et al.* RNA CoMPASS: a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PloS One* **9**, e89445 (2014).
45. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinforma. Oxf. Engl.* (2015). doi:10.1093/bioinformatics/btv639
46. Jørgensen, K. W., Rasmussen, M., Buus, S. & Nielsen, M. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* **141**, 18–26 (2014).
47. Larsen, M. V. *et al.* An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **35**, 2295–2303 (2005).
48. Nielsen, M., Lundegaard, C., Lund, O. & Keşmir, C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57**, 33–41 (2005).
49. Boisvert, F.-M. *et al.* A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. *Mol. Cell. Proteomics* **11**, M111.011429–M111.011429 (2012).
50. Duan, F. *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* **211**, 2231–2248 (2014).
51. Janeway's Immunobiology: 9780815345312: Medicine & Health Science Books @ Amazon.com. at <<http://www.amazon.com/Janeways-Immunobiology-Kenneth-Murphy/dp/0815345313>>
52. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
53. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014)
54. Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* 366, 1090–1098 (2012).
55. Hunt DF, Henderson RA, Shabanowitz J, Sakaguchi K, Michel H, Sevilir N, Cox AL, Appella E, Engelhard VH. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 1992. 255: 1261-1263.
56. Zarling AL, Polefrone JM, Evans AM, Mikesch LM, Shabanowitz J, Lewis ST, Engelhard VH, Hunt DF. Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy. *Proc Natl Acad Sci U S A.* 2006 Oct 3;103(40):14889-94.
57. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics.* 2015 Mar;14(3):658-73. doi: 10.1074/mcp.M114.042812.
58. Abelin JG, Trantham PD, Penny SA, Patterson AM, Ward ST, Hildebrand WH, Cobbold M, Bai DL, Shabanowitz J, Hunt DF. **Complementary IMAC enrichment**

**methods for HLA-associated phosphopeptide identification by mass spectrometry.** Nat Protoc. 2015 Sep;10(9):1308-18. doi: 10.1038/nprot.2015.086. Epub 2015 Aug 6

59. Barnstable CJ, Bodmer WF, Brown G, Galfre G, Milstein C, Williams AF, Ziegler A. Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. Cell. 1978 May;14(1):9-20.

60. Goldman JM, Hibbin J, Kearney L, Orchard K, Th'ng KH. HLA-DR monoclonal antibodies inhibit the proliferation of normal and chronic granulocytic leukaemia myeloid progenitor cells. Br J Haematol. 1982 Nov;52(3):411-20.

61. Eng JK, Jahan TA, Hoopmann MR. **Comet: an open-source MS/MS sequence database search tool.** Proteomics. 2013 Jan;13(1):22-4. doi: 10.1002/pmic.201200439. Epub 2012 Dec 4.

62. Eng JK, Hoopmann MR, Jahan TA, Egertson JD, Noble WS, MacCoss MJ. **A deeper look into Comet--implementation and features.** J Am Soc Mass Spectrom. 2015 Nov;26(11):1865-74. doi: 10.1007/s13361-015-1179-x. Epub 2015 Jun 27.

63. Lukas Käll, Jesse Canterbury, Jason Weston, William Stafford Noble and Michael J. MacCoss. **Semi-supervised learning for peptide identification from shotgun proteomics datasets.** Nature Methods 4:923 – 925, November 2007

64. Lukas Käll, John D. Storey, Michael J. MacCoss and William Stafford Noble. **Assigning confidence measures to peptides identified by tandem mass spectrometry.** Journal of Proteome Research, 7(1):29-34, January 2008

65. Lukas Käll, John D. Storey and William Stafford Noble. **Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry.** Bioinformatics, 24(16):i42-i48, August 2008

66. Bo Li and C. olin N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a referenfe genome. BMC Bioinformatics, 12:323, August 2011

67. Hillary Pearson, Tariq Daouda, Diana Paola Granados, Chantal Durette, Eric Bonneil, Mathieu Courcelles, Anja Rodenbrock, Jean-Philippe Laverdure, Caroline Côté, Sylvie Mader, Sébastien Lemieux, Pierre Thibault, and Claude Perreault. MHC class I-associated peptides derive from selective regions of the human genome. The Journal of Clinical Investigation, 2016,

68. Juliane Liepe, Fabio Marino, John Sidney, Anita Jeko, Daniel E. Bunting, Alessandro Sette, Peter M. Klotzel, Michael P. H. Stumpf, Albert J. R. Heck, Michele Mishto. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. Science, 21, October 2016.

69. Mommen GP., Marino, F., Meiring HD., Poelen, MC., van Gaans-van den Brink, JA., Mohammed S., Heck AJ., and van Els CA. Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. Mol Cell Proteomics 15(4): 1412-1423, April 2016.

70. Sebastian Kreiter, Mathias Vormehr, Niels van de Roemer, Mustafa Diken, Martin Löwer, Jan Diekmann, Sebastian Boegel, Barbara Schrörs, Fulvia Vascotto, John C. Castle, Arbel D. Tadmor, Stephen P. Schoenberger, Christoph Huber, Özlem Türeci, and Ugur Sahin. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. Nature 520, 692-696, April 2015.71. Tran E., Turcotte S., Gros A., Robbins P.F., Lu Y.C., Dudley M.E., Wunderlich J.R., Somerville R.P., Hogan K., Hinrichs C.S., Parkhurst M.R., Yang J.C., Rosenberg S.A. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science 344(6184) 641-645, May 2014.72. Andreatta M., Karosiene E., Rasmussen M., Stryhn A., Buus S., Nielsen M. Accurate pan-specific

prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* 67(11-12) 641-650, November 2015.

73. Nielsen, M., Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 10:296, September 2009.

74. Nielsen, M., Lundegaard, C., Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8:238, July 2007.

75. Zhang, J., et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics*. 11(4):1-8. 1/2/2012.

**CLAIMS**

1. A method of identifying a subset of patients for treatment, comprising:  
obtaining, for each patient, at least one of exome, transcriptome, or whole genome tumor nucleotide sequencing data from tumor cells and normal cells of the patient, wherein the tumor nucleotide sequencing data is used to obtain 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 for the patient comprises at least one alteration that makes it distinct from a corresponding wild-type parental peptide sequence identified from the normal cells of the patient;  
generating, for each patient, a set of numerical presentation likelihoods for the set of neoantigens for the patient by inputting the peptide sequences of each of the set of neoantigens into a machine-learned presentation model, each presentation likelihood representing the likelihood that a corresponding neoantigen is presented by one or more MHC alleles on the surface of the tumor cells of the patient, the set of presentation likelihoods having been identified at least based on mass spectrometry data;  
identifying, for each patient, one or more neoantigens from the set of neoantigens for the patient;  
determining, for each patient, a utility score indicating an estimated number of neoantigens presented on the surface of the tumor cells of the patient as determined by the corresponding presentation likelihoods for the one or more neoantigens for the patient; and  
selecting the subset of patients for treatment, each patient in the subset of patients associated with a utility score that satisfies a predetermined inclusion criteria.
2. The method of claim 1, wherein identifying the one or more neoantigens for the patient comprises selecting a subset of neoantigens in the set of neoantigens for the patient.
3. The method of claim 2, wherein the subset of neoantigens are neoantigens having highest presentation likelihoods in the set of presentation likelihoods for the patient.
4. The method of claim 1, further comprising treating, for each patient in the selected

subset of patients, with a corresponding neoantigen vaccine including at least one of the one or more neoantigens identified for the patient.

5. The method of claim 1, further comprising identifying, for each patient in the selected subset of patients, one or more T-cells or T-cell receptors that are antigen-specific for at least one of the one or more neoantigens identified for the patient.
6. The method of claim 1, wherein identifying the one or more neoantigens for the patient comprises selecting the entire set of neoantigens identified for the patient.
7. The method of claim 6, further comprising administering checkpoint inhibitor therapy to each patient in the selected subset of patients.
8. The method of claim 1, wherein selecting the subset of patients for treatment comprises selecting the subset of patients having tumor mutation burden (TMB) above a minimum threshold, wherein the TMB for a patient indicates a number of neoantigens in the set of neoantigens associated with the patient.
9. The method of claim 1, wherein selecting the subset of patients for treatment comprises selecting the subset of patients having utility scores above a minimum threshold.
10. The method of claim 1, wherein the utility score is a summation of the presentation likelihoods for each neoantigen in the identified subset of neoantigens of the patient.
11. The method of claim 1, wherein the utility score is a probability that a number of presented neoantigens in the identified one or more neoantigens for the patient is above a minimum threshold.
12. The method of claim 1, wherein the machine-learned presentation model comprises:  
a plurality of parameters identified at least based on a training data set comprising:  
labels obtained by mass spectrometry measuring presence of peptides bound to  
at least one MHC allele identified as present in at least one of a plurality of  
samples,  
training peptide sequences including information regarding a plurality of  
amino acids that make up the training peptide sequences and a set of  
positions of the amino acids in the training peptide sequences, and  
at least one MHC allele associated with the training peptide sequences; and

a function representing a relation between the peptide sequences and the presentation likelihoods based on the plurality of parameters.

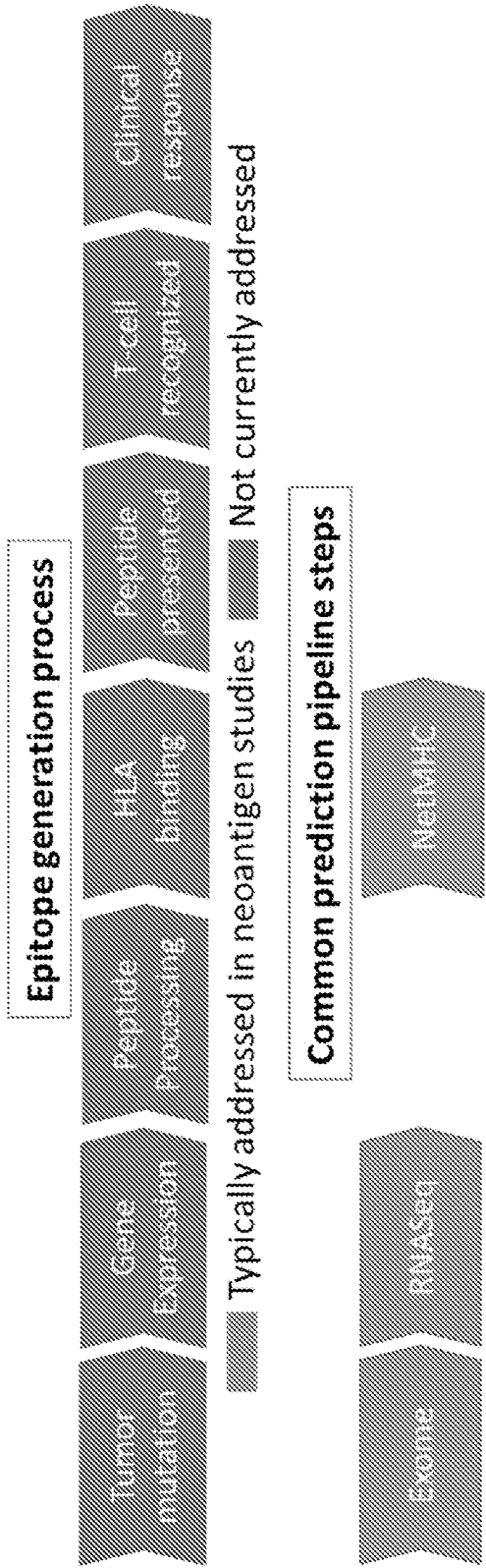
13. The method of claim 12, wherein the training data set further comprises at least one of:
  - (a) data associated with peptide-MHC binding affinity measurements for at least one of the isolated peptides; and
  - (b) data associated with peptide-MHC binding stability measurements for at least one of the isolated peptides.
14. The method of claim 1, 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.
15. The method of claim 1, 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.
16. The method of claim 1, 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.
17. The method of claim 1, wherein inputting the peptide sequences into the machine-learned presentation model comprises:

applying the machine-learned presentation model to the peptide sequence of each neoantigen to generate a dependency score for each of the one or more MHC alleles indicating whether a MHC allele will present the neoantigen based on the particular amino acids at the particular positions of the peptide sequence.
18. The method of claim 17, wherein inputting the peptide sequences into the machine-learned presentation model 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.

19. The method of claim 18, wherein transforming the dependency scores models the presentation of the neoantigen as mutually exclusive across the one or more class MHC alleles.
20. The method of claim 17, wherein inputting the peptide sequences into the machine-learned presentation model 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.

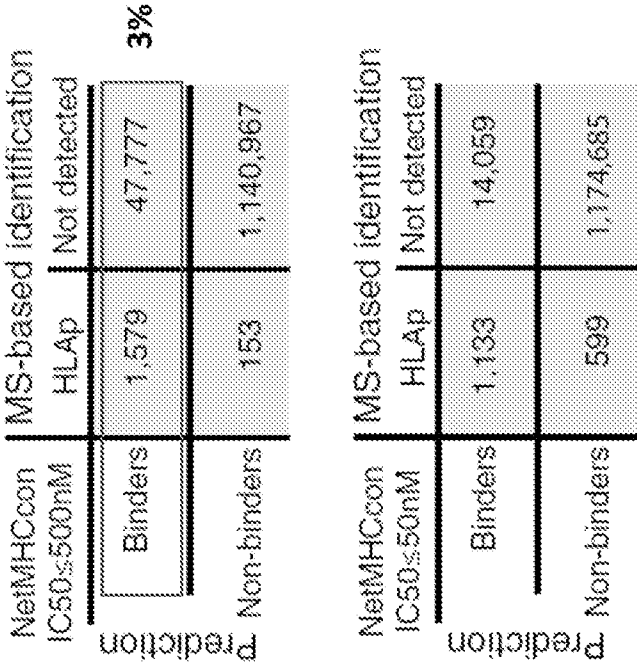
# Current clinical approaches to neoantigen identification



- Epitope generation process is typically modeled incompletely
- Binding generally considered most selective step and is included, but aggregate impact of other steps may be large
- Using expression and binding filters alone has potential for high false positive rate

FIG. 1A

Most recent literature suggests <5% of predicted bound peptides can be found presented on tumor cells



- Only 7/170 (4%) predicted neo-epitopes were eluted and detected by mass-spec
- Possibly due to sensitivity limits of particular mass-spec detection approach

FIG. 1B

Yadav, Nature Vol 515 17 November 2014

Bassani-Sternberg M, Mol Cell Proteomics. 2015;14(3):658-73

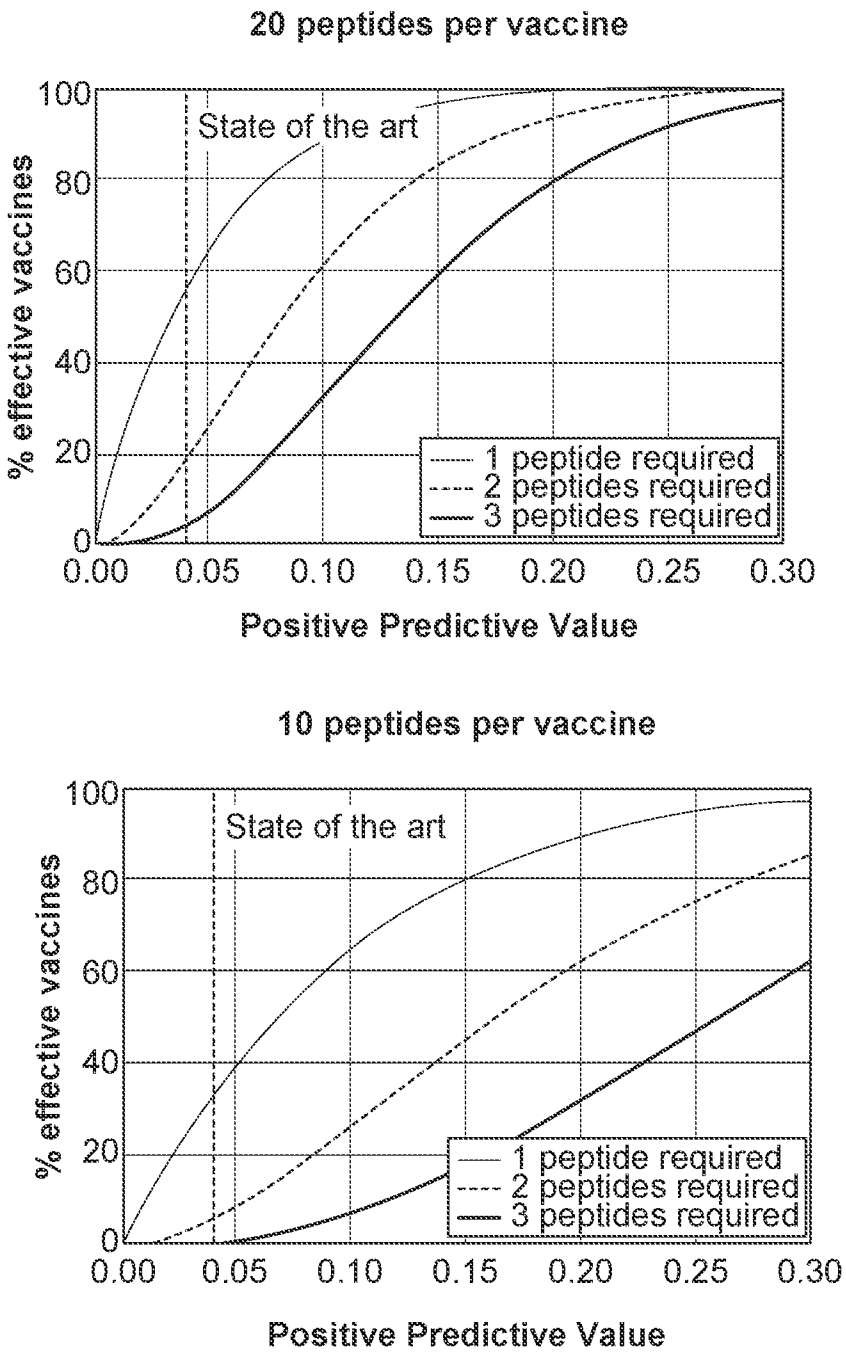


**FIG. 1C**

Binding prediction necessary but not sufficient

Binding affinity prediction vs. mass-spec peptide detection in JY (EBV)-immortalized LCL

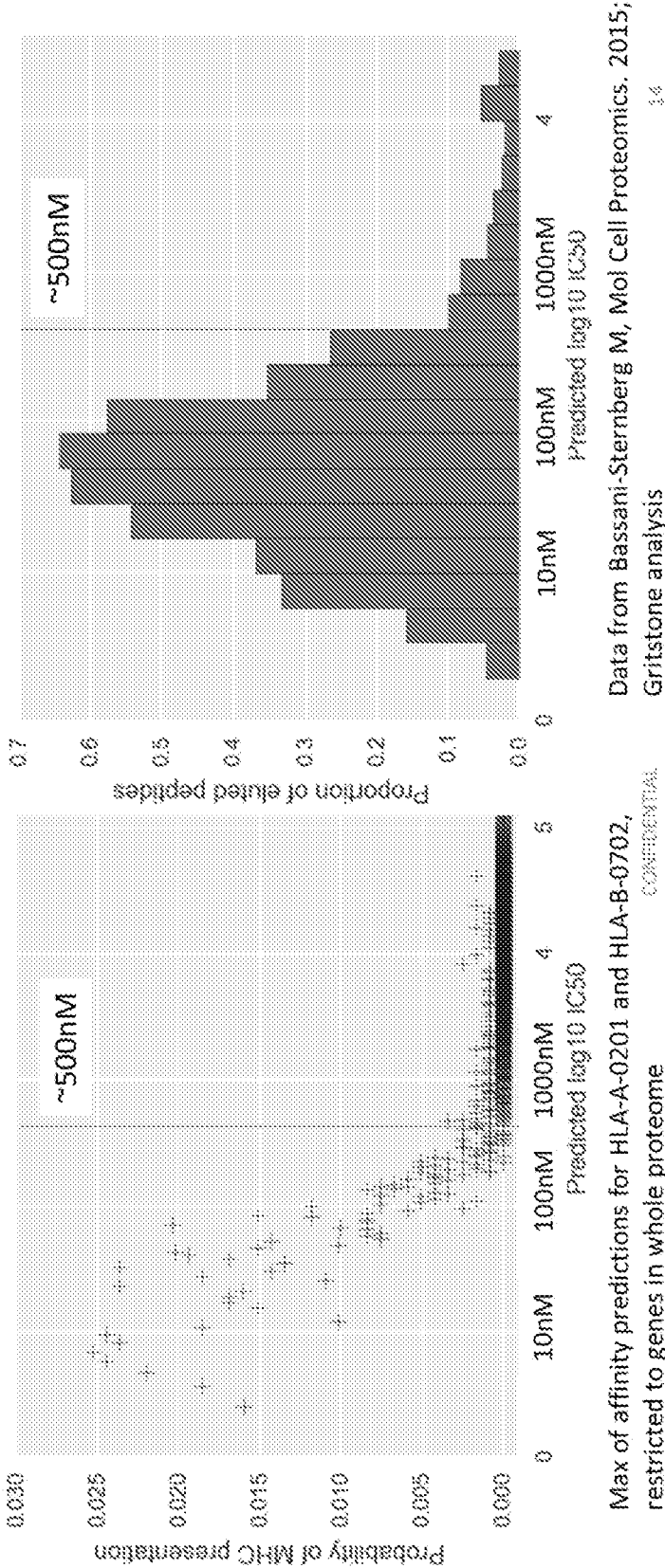


FIG. 1D

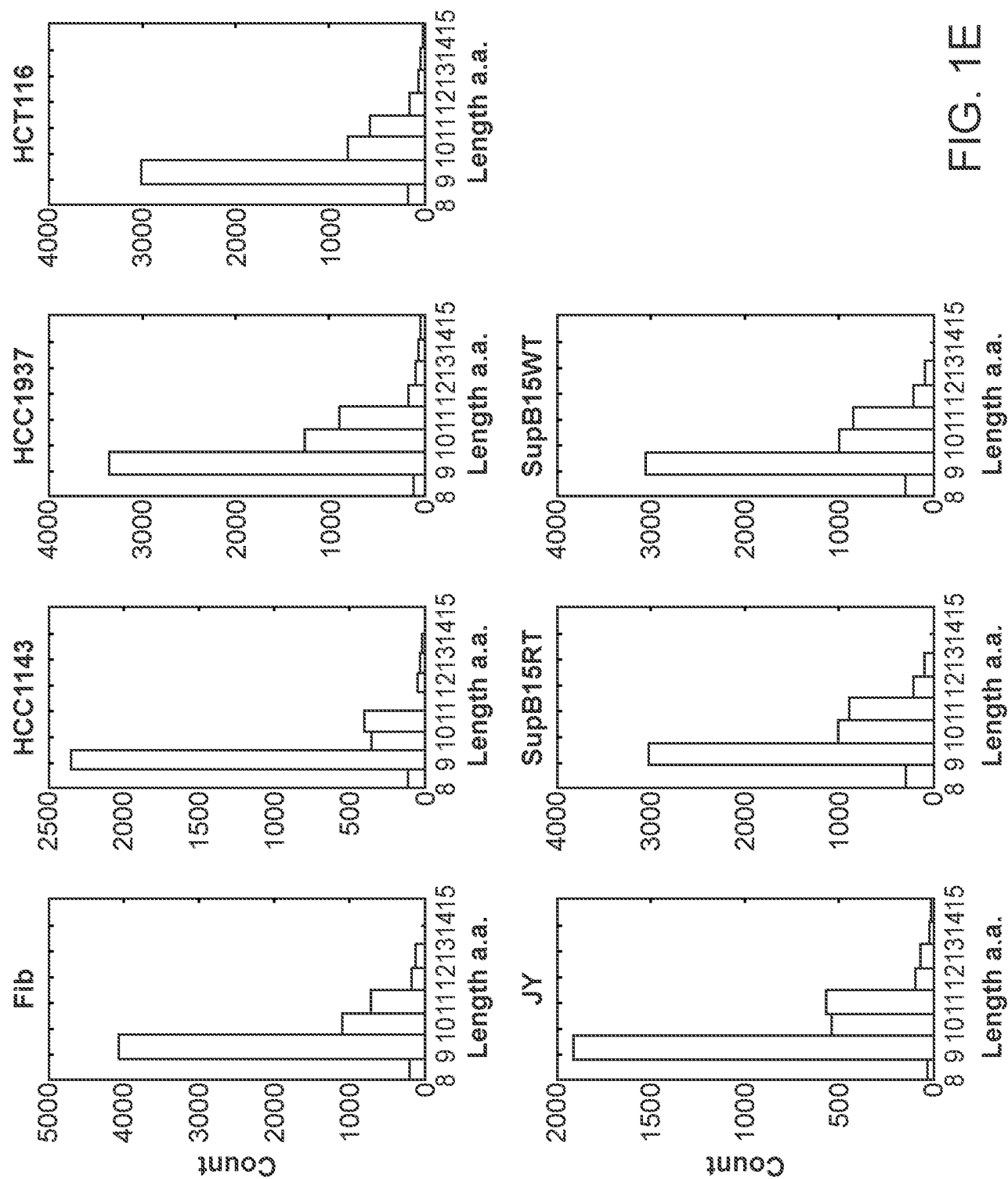
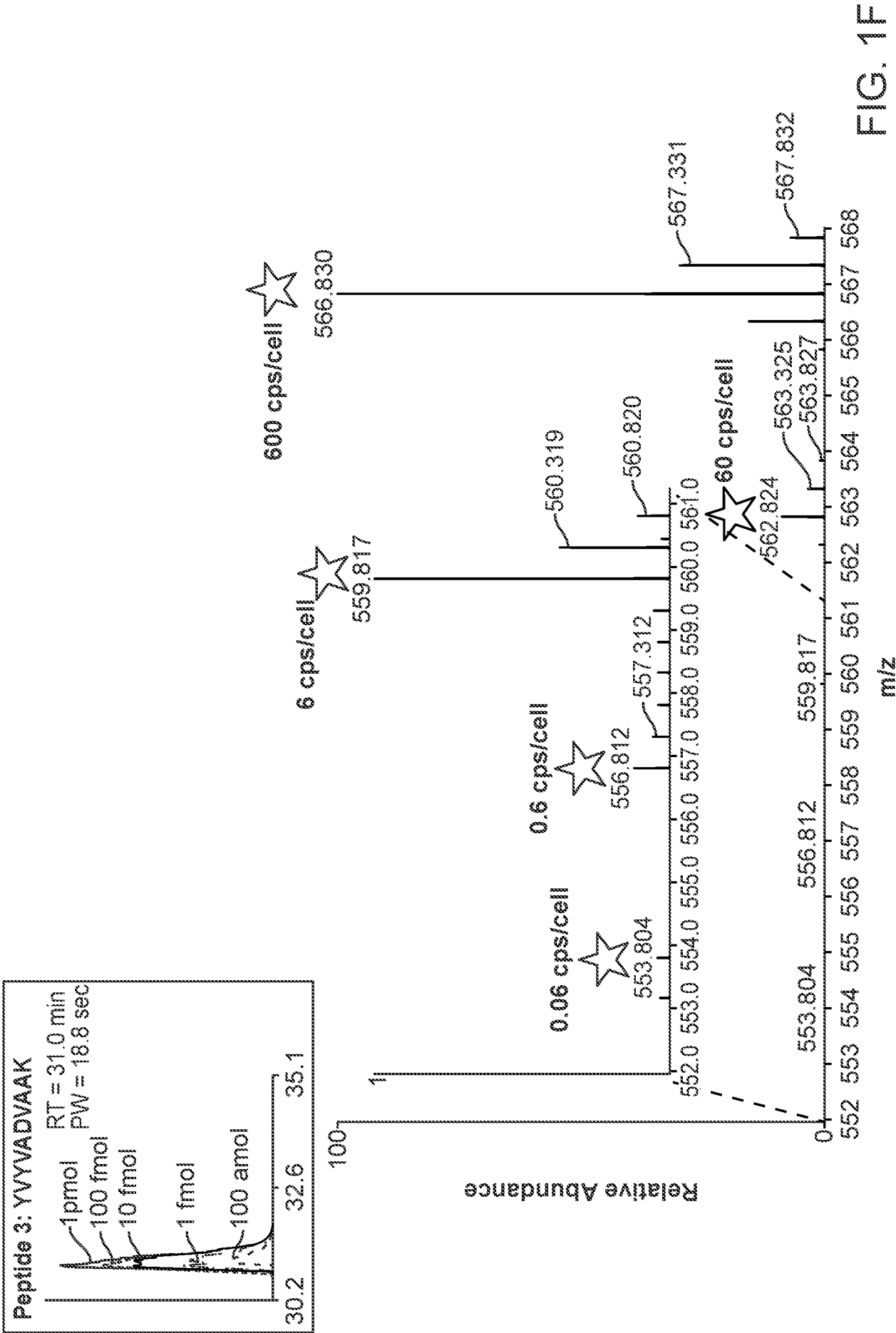


FIG. 1E



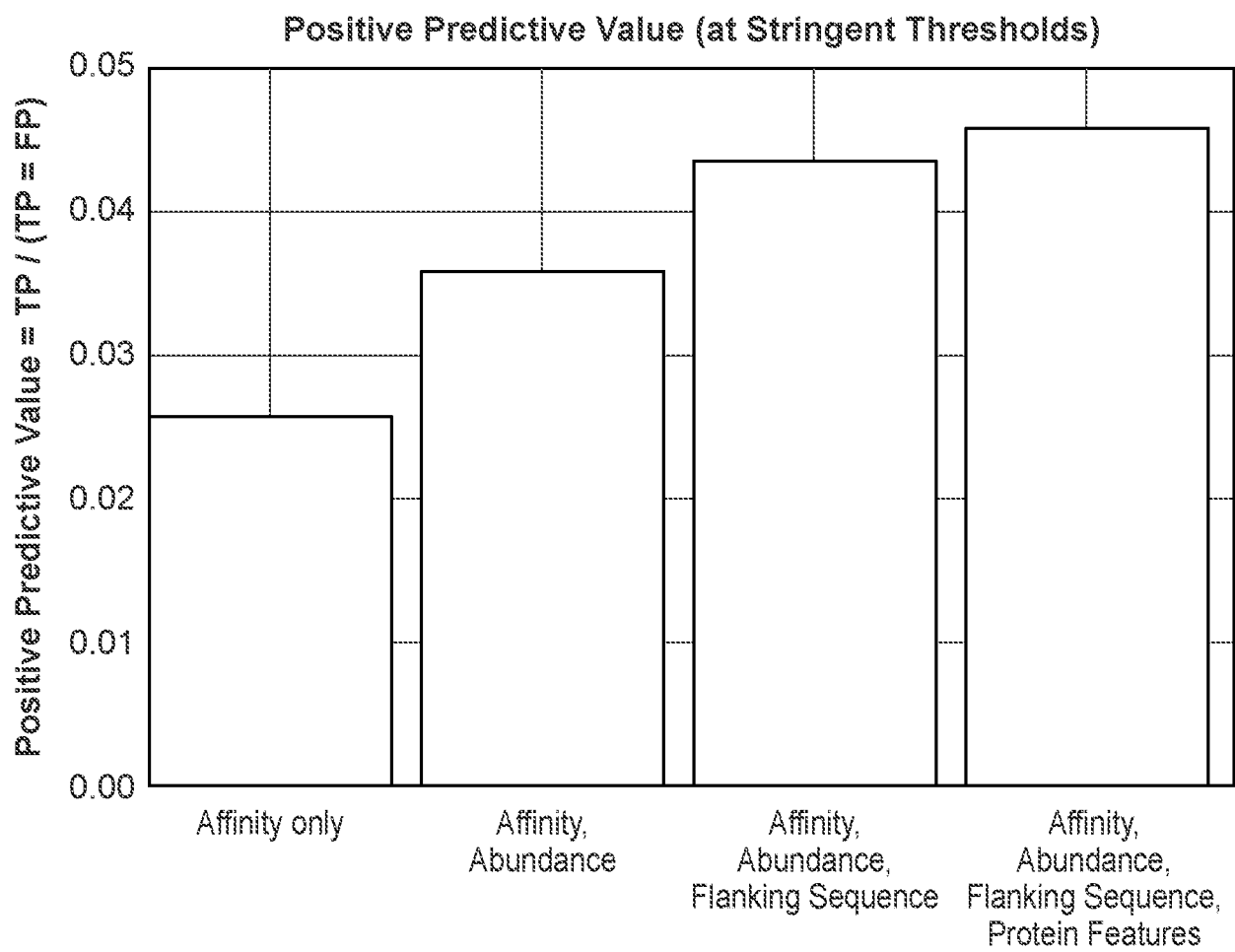


FIG. 1G

8/26

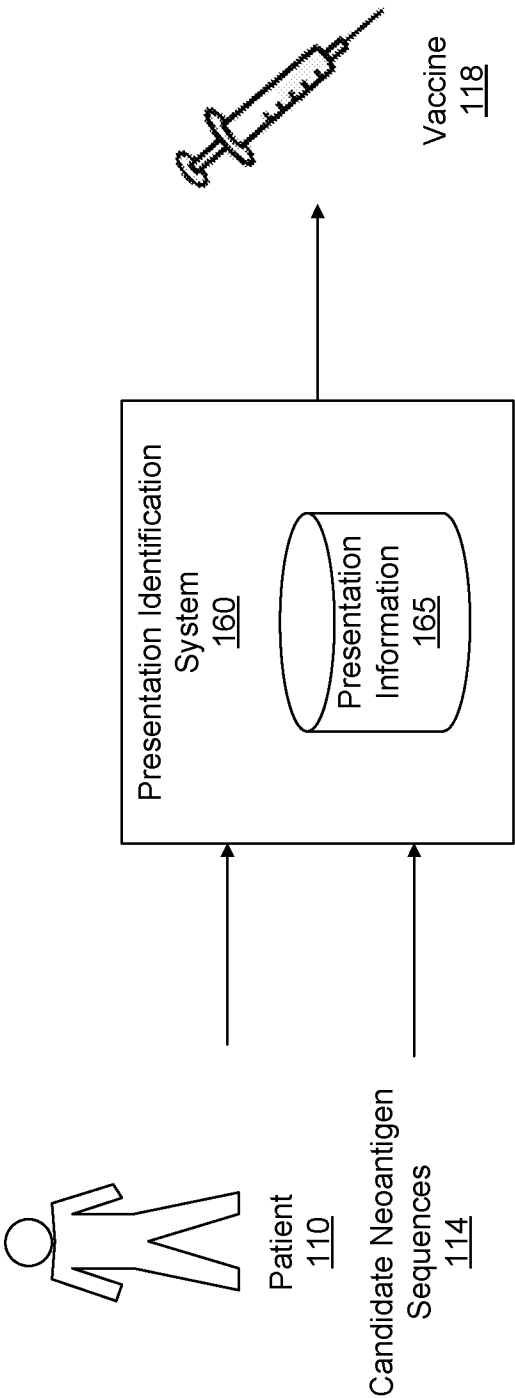


FIG. 2A

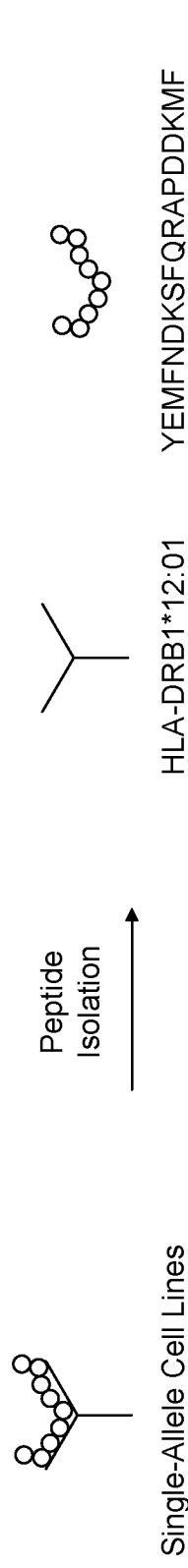


FIG. 2B

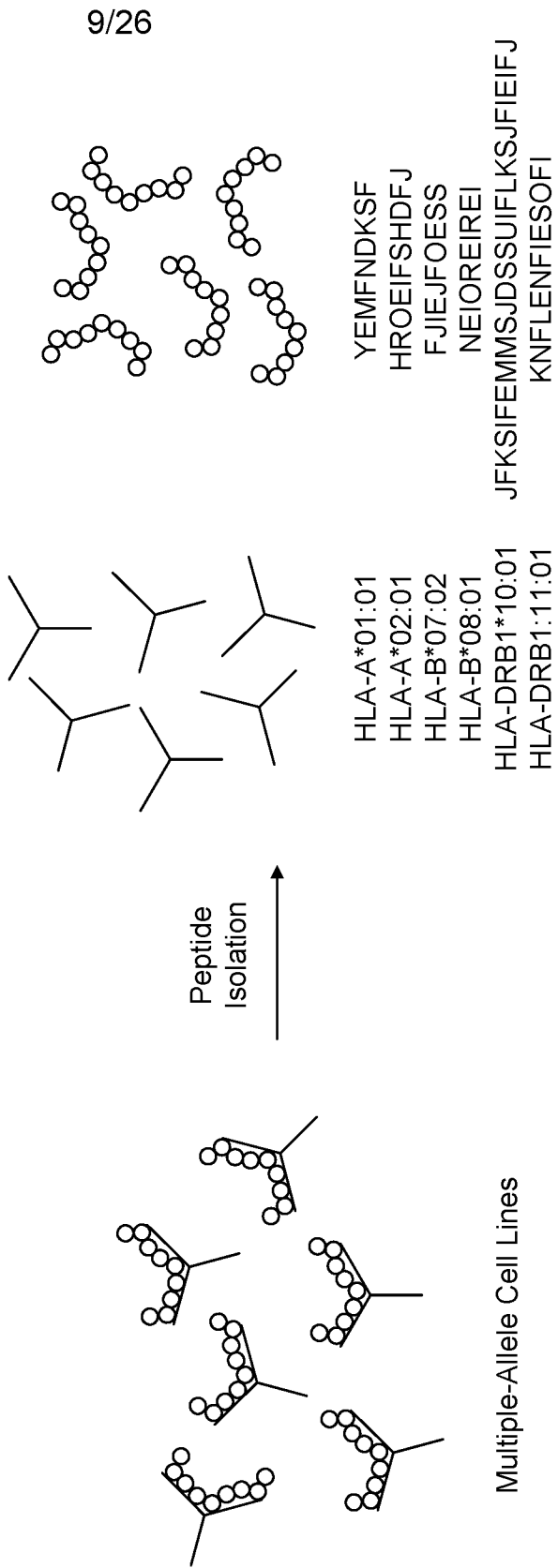


FIG. 2C

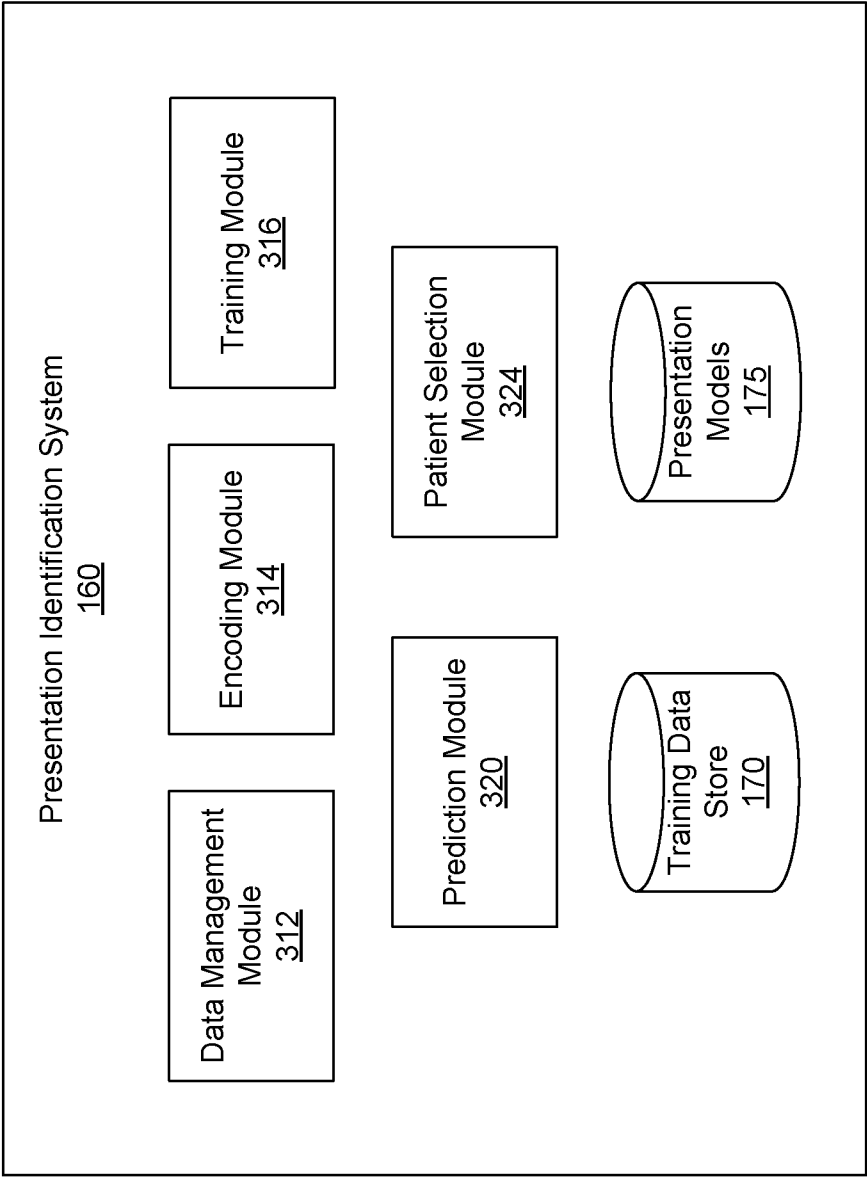


FIG. 3

Training Data  
170A

Allele-Dependent ( $\mathbf{x}^i$ )

Allele-Independent ( $\mathbf{w}^i$ )

| 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 <sup>2</sup>       | Not Presented   |
| FIEUHFWI                   | 1500                  | 15                    | HLA-C*01:03      | FEGRKUOOI                     | 10 <sup>-3</sup>      | Presented       |
| FEWRHRJTRUJR               | 650                   | 20                    | HLA-C*01:03      | PJFIOEJOIJGEIO                | 10 <sup>1</sup>       | Presented       |
| QIEJOEIJIE                 | 500                   | 1                     | HLA-B*07:02      | PJFIOEJOIJGEIO                | 1                     | Presented       |
|                            | 600                   | 14                    | HLA-C*01:03      |                               |                       |                 |
|                            | 1200                  | 7                     | HLA-A*01:01      |                               |                       |                 |

11/26

FIG. 4

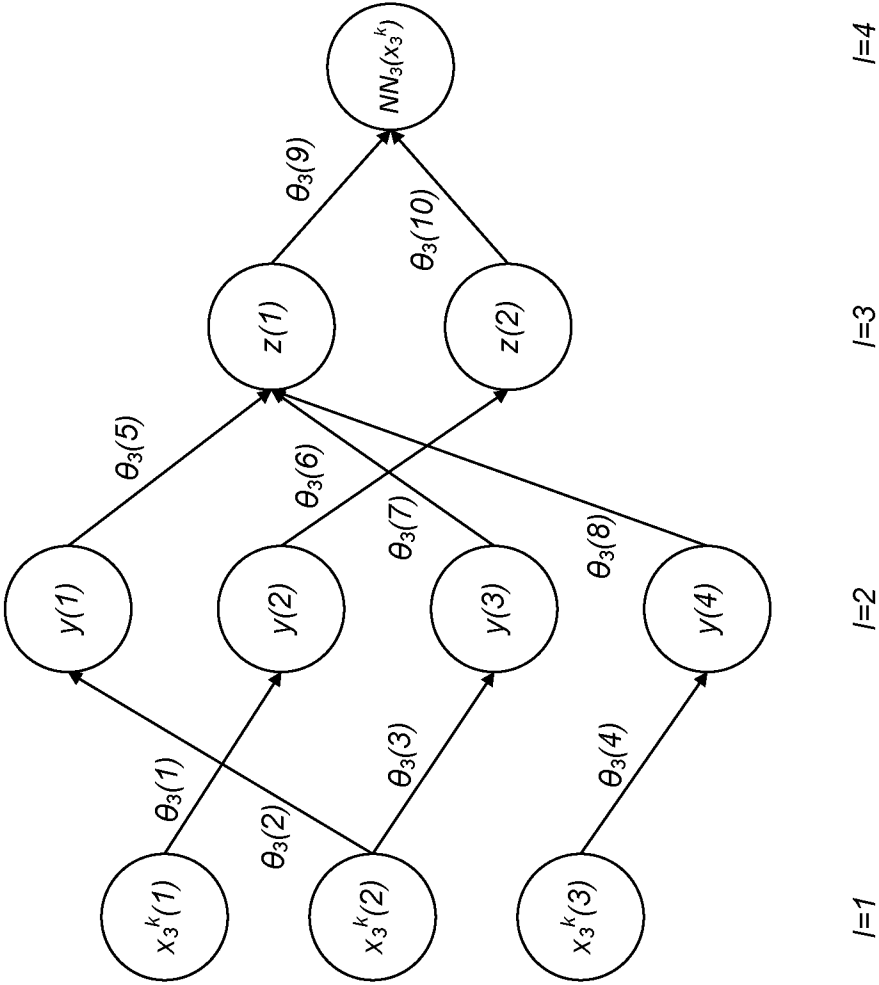


FIG. 5

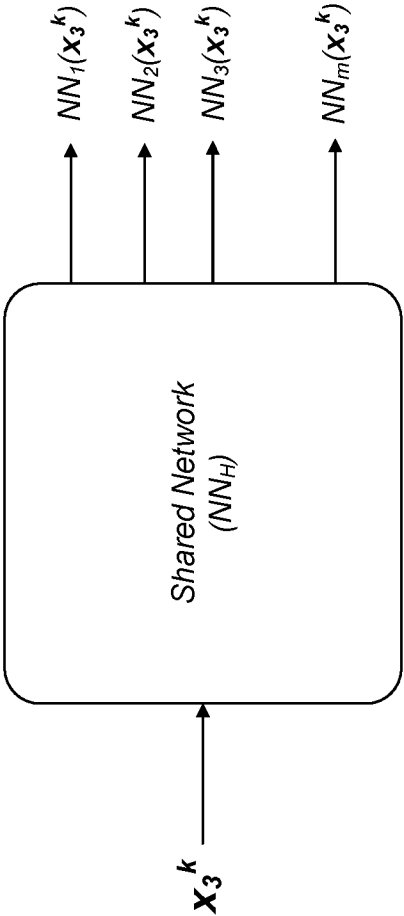


FIG. 6A

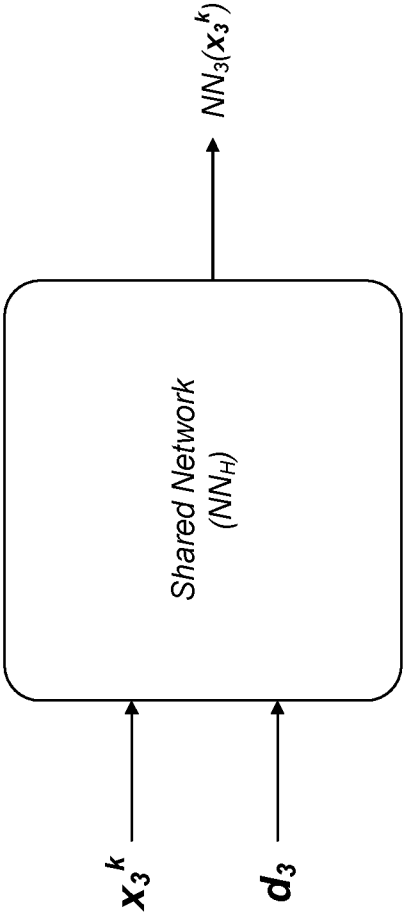


FIG. 6B

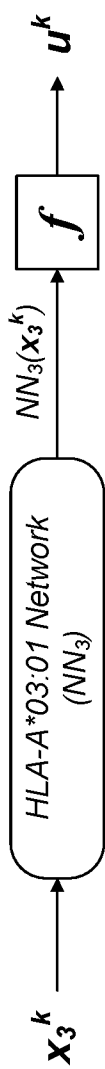


FIG. 7

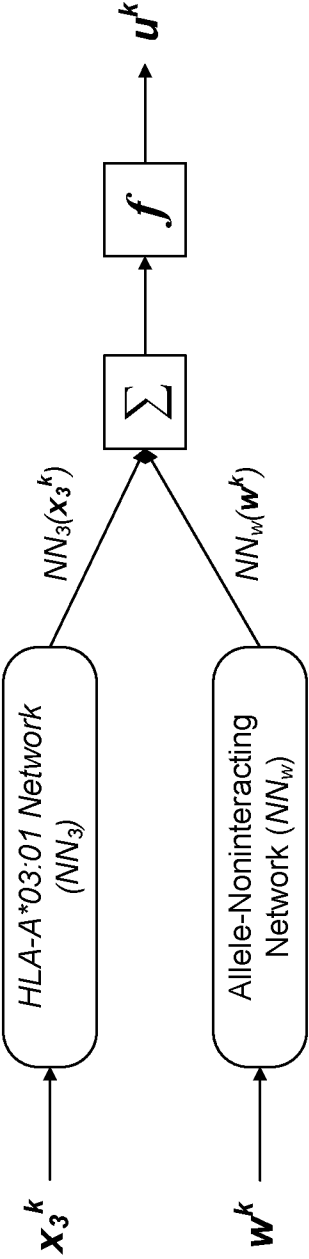


FIG. 8

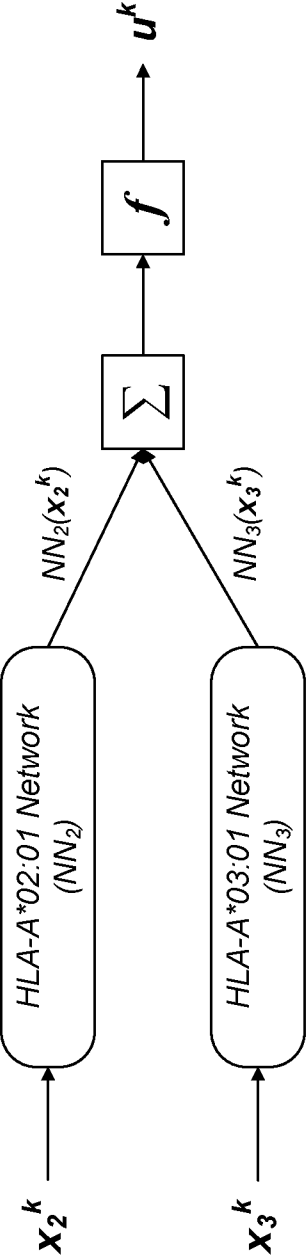


FIG. 9

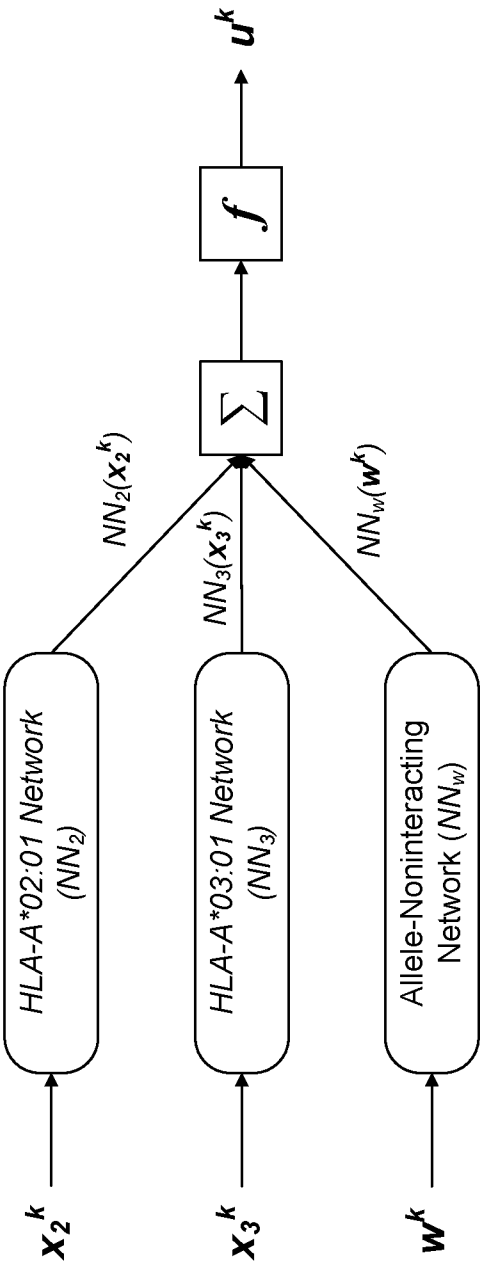


FIG. 10

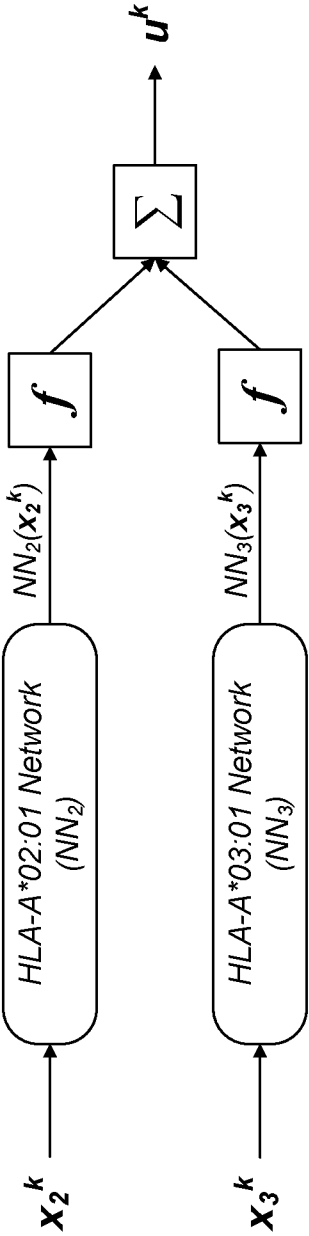


FIG. 11

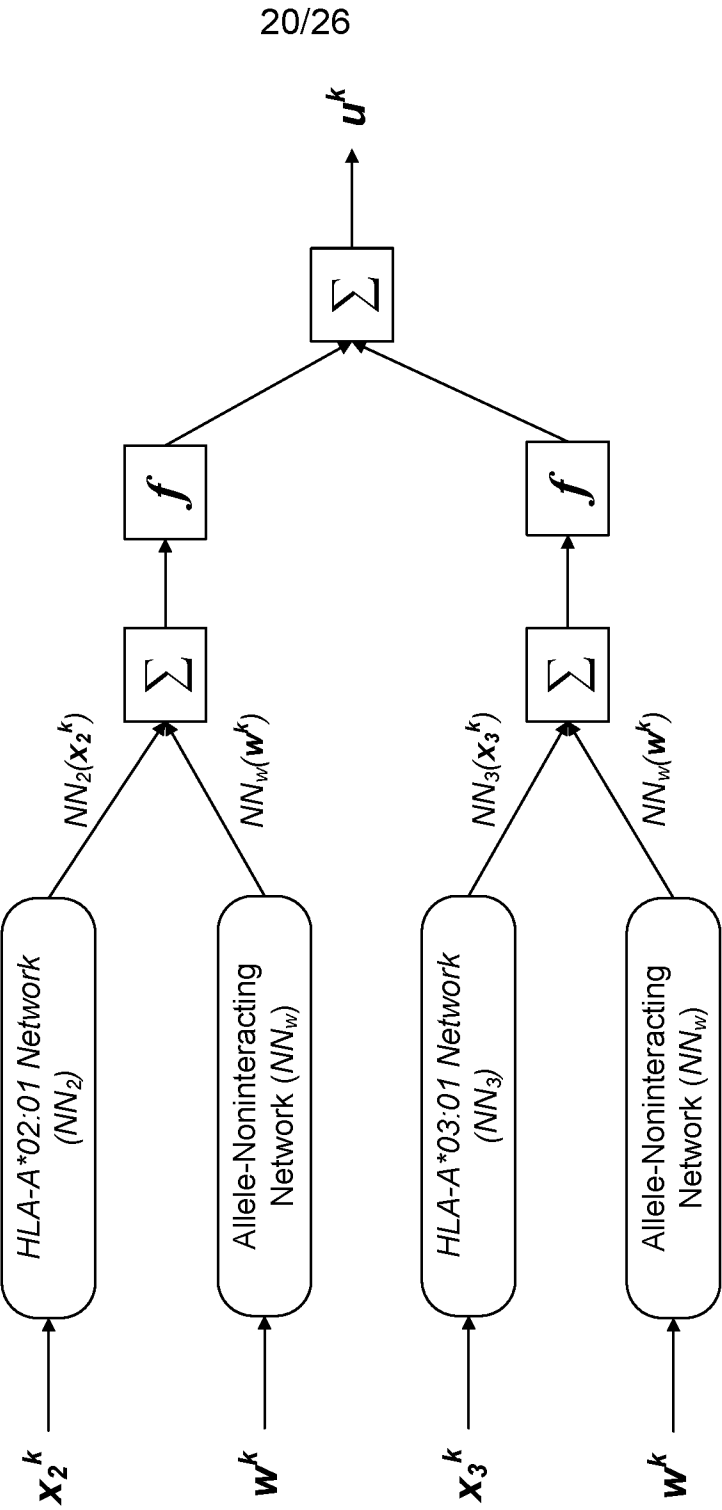


FIG. 12

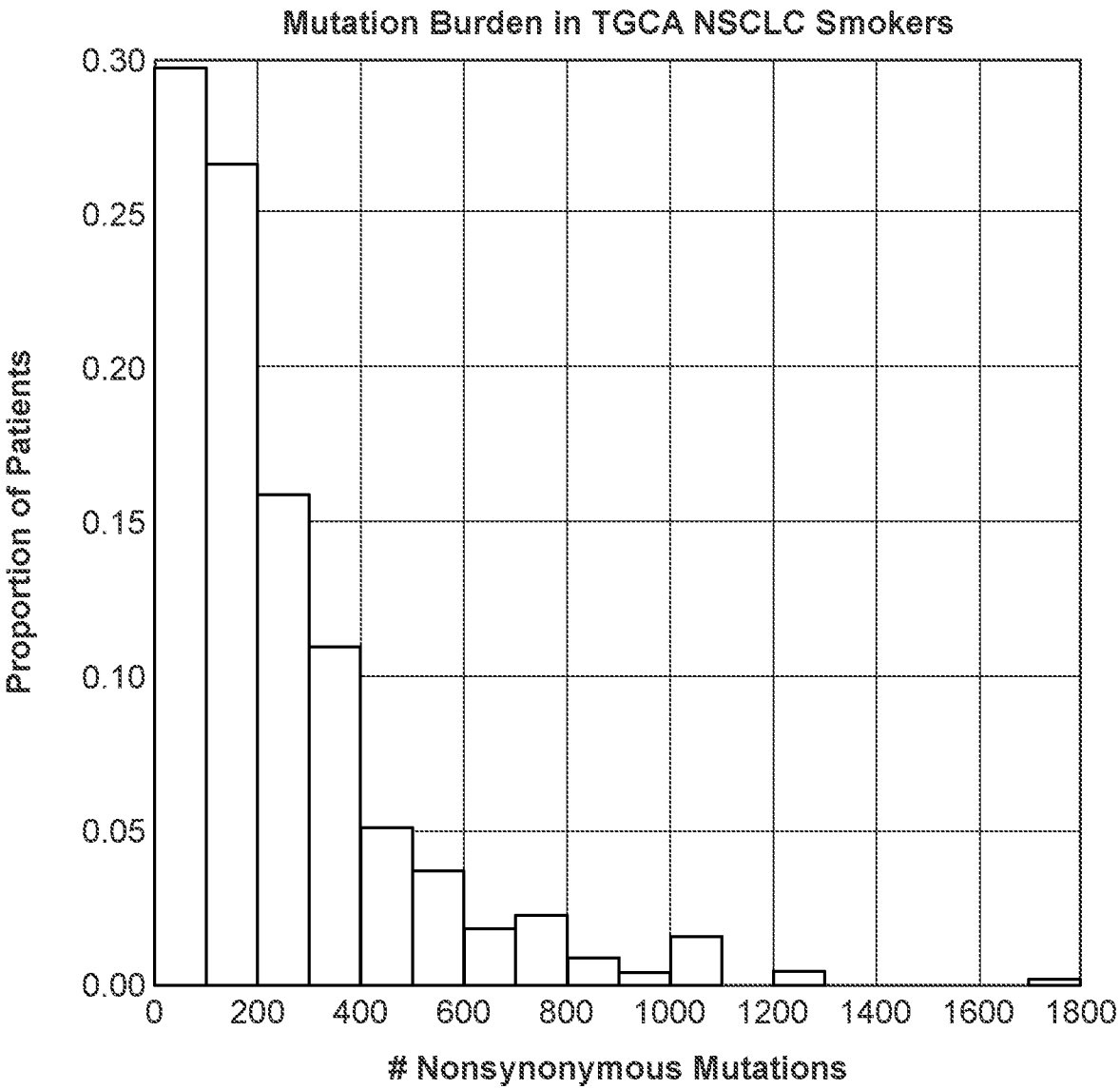


FIG. 13A

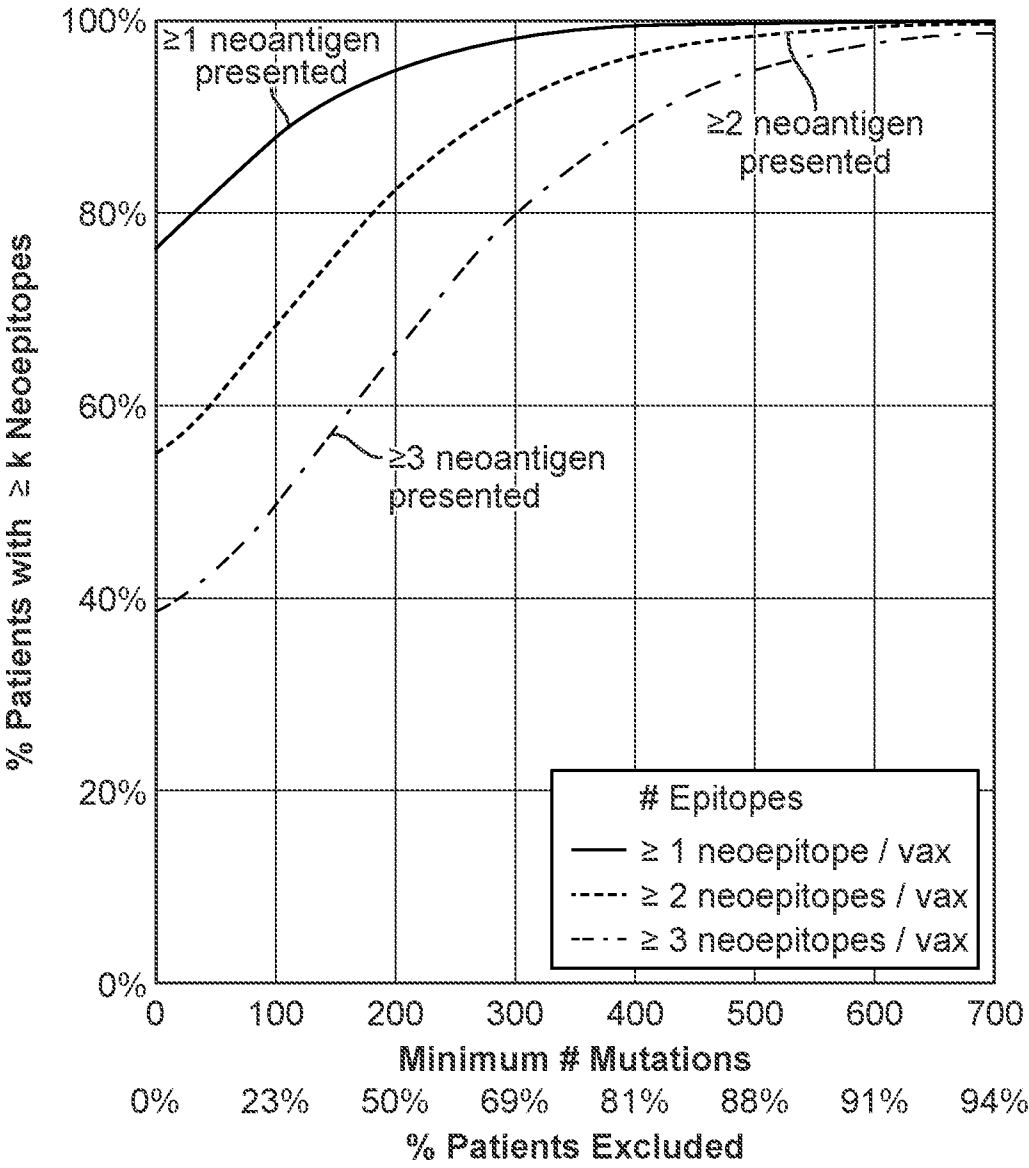


FIG. 13B

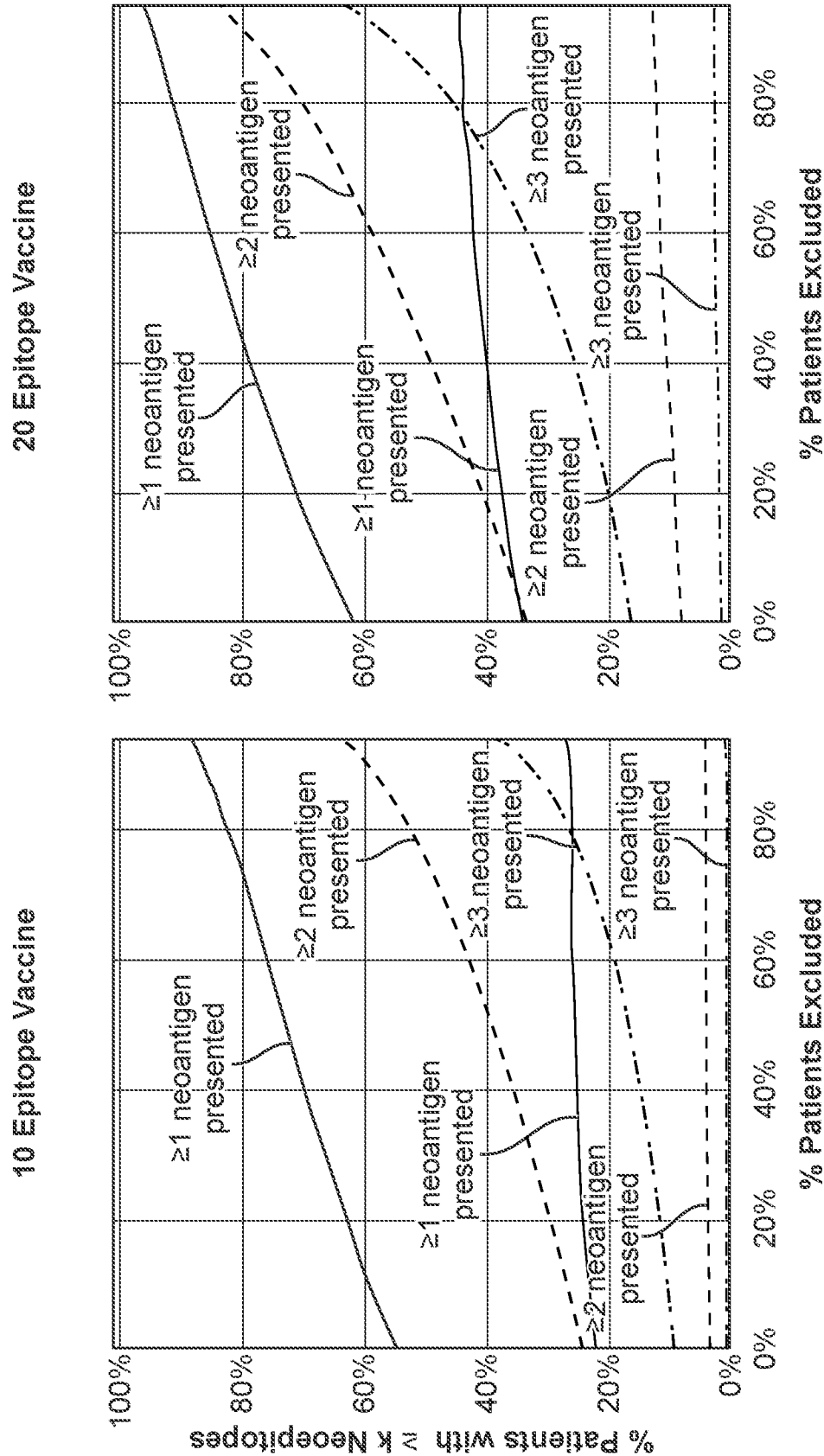


FIG. 13C

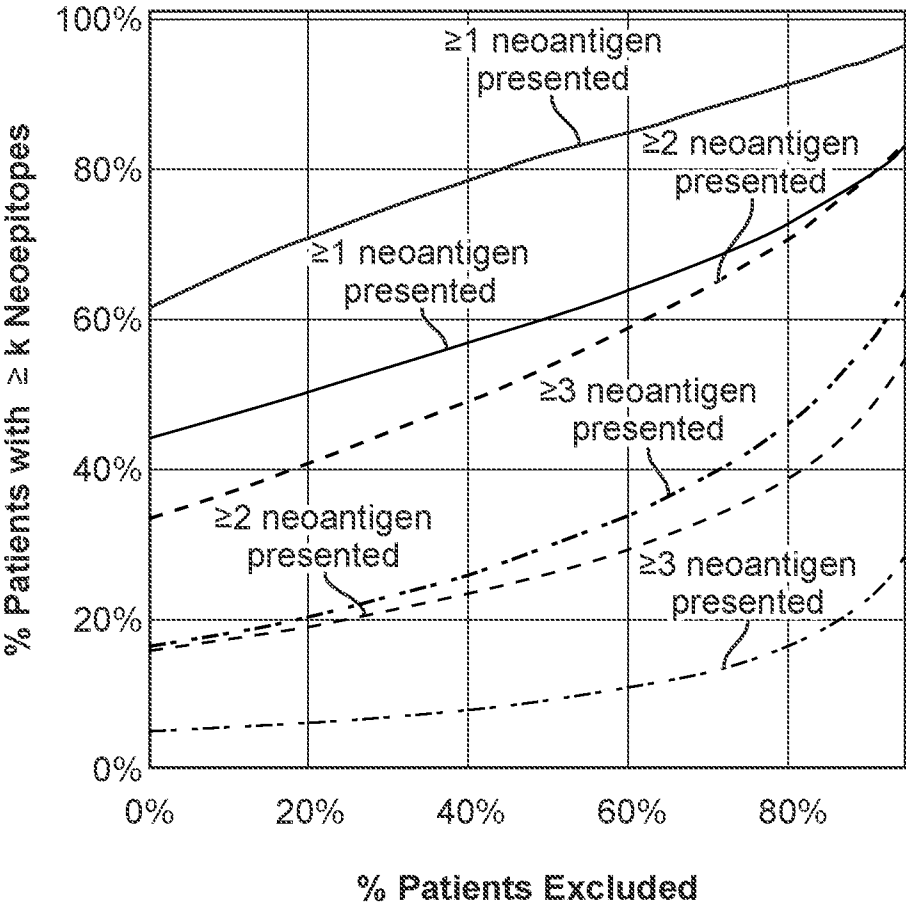


FIG. 13D

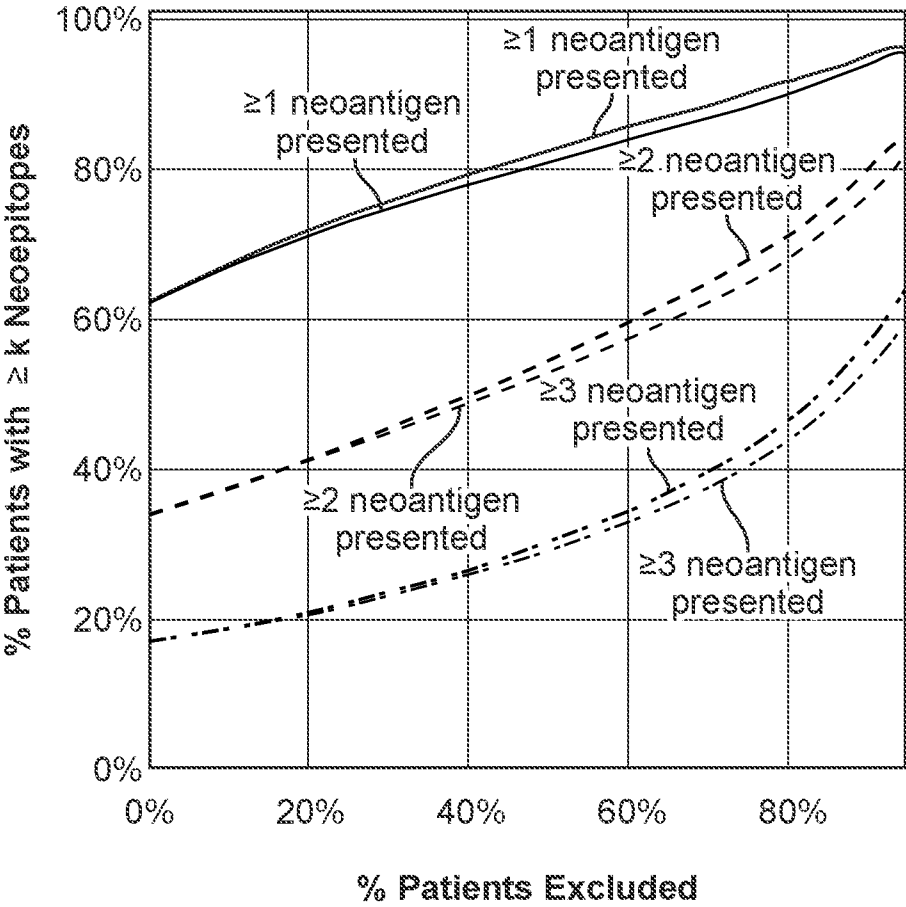


FIG. 13E

26/26

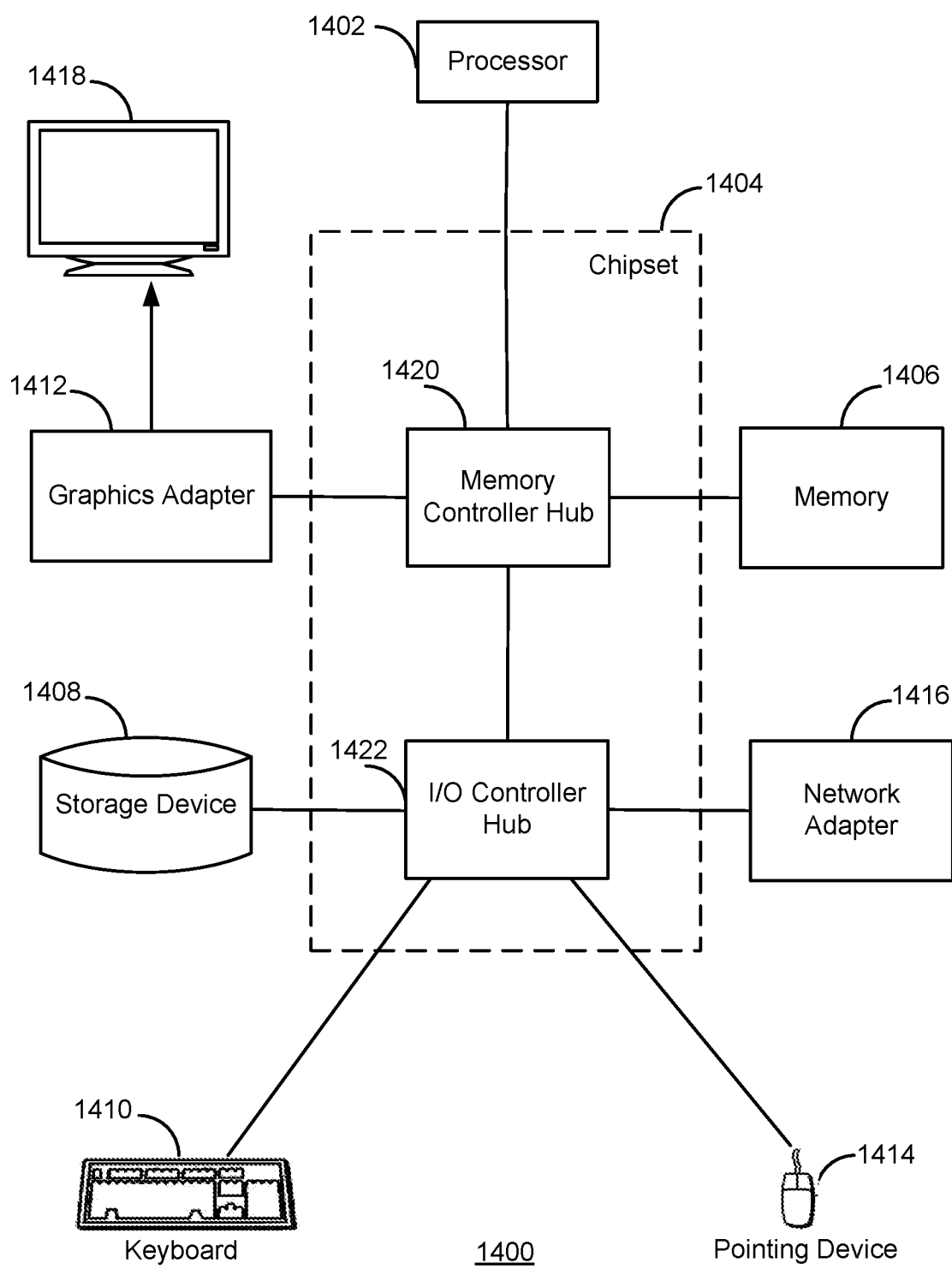


FIG. 14

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 18/36571

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 19/22; A61K 39/00, 38/10, 35/15; C07K 7/08 (2018.01)

CPC - G06F 19/22; A61K 39/0011, 38/10, 35/15, 2039/70; C07K 14/4748

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category*       | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                         | Relevant to claim No.                 |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| X<br>-----<br>Y | US 2016/0101170 A1 (GENERAL HOSPITAL CORPORATION) 14 April 2016 (14.04.2016). Especially para [0008], [0009], [0090], [0093], [0095], [0101], [0104], [0133], [0142], [0280], [0398], [[0406-0408], [0447], [0450], [0459], [0466], [0478].                                                | 1-7, 9-11, 15-20<br>-----<br>8, 12-14 |
| Y               | MILLER et al. Correlation Between Somatic Mutation Burden, Neoantigen Load and Progression Free Survival in Multiple Myeloma: Analysis of MMRF CoMMpass Study. Blood December 2016 Vol 128 No 22 Pages 193. Especially abstract.                                                           | 8                                     |
| Y               | ANDREATTA et al. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics 2015 November Vol 67 Pages 641-650 [in NIH Author's manuscript version, pp 1-18]. Especially pg 3 para 5, pg 4 para 3-4, pg 8 para 6. | 12-14                                 |
| A               | YADAV et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature 27 November 2014 Vol 515 No 7528 Pages 572-576. Especially abstract, pg 573 fig 1A, pg 573 table 1.                                                                       | 1                                     |

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search

7 September 2018

Date of mailing of the international search report

03 OCT 2018

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents  
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774