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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 January 2019 (03.01.2019)

W!P O ¦ PCT

(10) International Publication Number
WO 2019/006022 Al

(51) International Patent Classification:
*G06F 19/22* (201 1.01) *A61K 38/10* (2006.01)
*A61K 39/00* (2006.01) *A61K 35/15* (2015.01)
*A61K 39/02* (2006.01) *C07K 7/08* (2006.01)

(21) International Application Number:
PCT/US20 18/039843

(22) International Filing Date:
27 June 2018 (27.06.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/525,673 27 June 2017 (27.06.2017) US

(71) Applicants: **THE BROAD INSTITUTE, INC.** [US/US]; 415 Main Street, Cambridge, Massachusetts 02142 (US). **THE GENERAL HOSPITAL CORPORATION** [US/US]; 55 Fruit Street, Boston, Massachusetts 021 14 (US).

(72) Inventor; and
(71) Applicant: **XAVIER, Ramnik** [US/US]; c/o Fruit Street, Boston, Massachusetts 021 14 (US).

(72) Inventors: **GRAHAM, Daniel B.;** c/o 415 Main Street, Cambridge, Massachusetts 02142 (US). **LUO, Chengwei;** c/o 415 Main Street, Cambridge, Massachusetts 02142 (US).

(74) Agent: **NIX, F. Brent** et al.; Johnson, Marcou & Isaacs, LLC, P.O. Box 69 1, Hoschton, Georgia 30548 (US).

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

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

WO 2019/006022 A1

**(54) Title:** SYSTEMS AND METHODS FOR MHC CLASS II EPITOPE PREDICTION

**(57) Abstract:** A system and method for prediction of immunodominant epitopes is provided herein. MHCII peptidomics was used to discover complex bacterial epitopes and host antigen processing pathways. Novel insights into the features of antigenicity are leveraged to build an algorithm for prediction of immunodominant epitopes. Use of immunodominant epitopes is described.

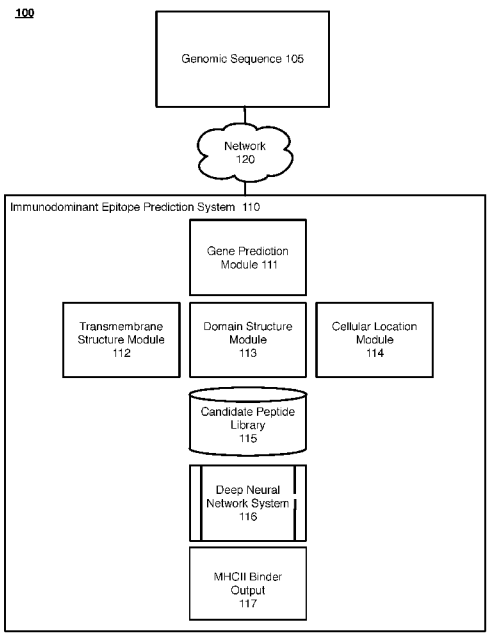



FIG. 1

*[Continued on next page]*

**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))*
— *with sequence listing part of description (Rule 5.2(a))*

# SYSTEMS AND METHODS FOR MHC CLASS II EPITOPE PREDICTION

## CROSS-REFERENCE TO RELATED APPLICATIONS

**[0001]** This application claims the benefit of U.S. Provisional Application No. 62/525,673, filed June 27, 2017. The entire contents of the above-identified applications are hereby fully incorporated herein by reference.

## STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

**[0002]** This invention was made with government support under grant numbers DK04335 1 and AI109725 awarded by the National Institutes of Health. The government has certain rights in the invention.

## TECHNICAL FIELD

**[0003]** This disclosure relates generally to methods and systems for the prediction of MHCII immunodominant epitopes requiring only an annotated genome sequence as an input and methods of using the predicted epitopes.

## BACKGROUND

**[0004]** Canonical CD4 T cell responses are restricted to cognate peptide antigen presented by MHCII, which is in turn dictated by sequence-specific interactions between the peptide backbone and MHCII binding groove (Babbitt et al., 1985; Stern et al., 1994). Degeneracy in this interaction allows for presentation of a broad spectrum of peptide antigens and promotes diverse responses to potential antigens. However, CD4 T cell responses tend to be constrained to a limited set of immunodominant epitopes even when the pool of available peptide epitopes is not limiting. Thus, the T cell response must be balanced with respect to the magnitude of the response against any given epitope (to maximize efficacy) and diversity towards multiple epitopes (to combat epitope escape). Despite intensive investigation, the factors controlling immunodominance and antigenicity are incompletely understood.

**[0005]** At the level of antigen presenting cells, several lysosomal pathways contribute to antigen processing and epitope selection (Kim and Sadegh-Nasseri, 2015). In this context, numerous thiol reductases (Arunachalam et al., 2000; Cresswell et al., 1999) and proteases (Hsieh et al., 2002; Hsing and Rudensky, 2005) perform redundant functions in converting native proteins into MHCII ligands. Sequence- and structure-specific preferences of these

enzymes for their substrates can bias which peptides are ultimately processed and loaded onto MHCII. Furthermore, kinetic parameters of the peptide- MHCII interaction impact the stability of the complex such that weak-binding peptides are replaced by means of HLA-DM editing (Miyazaki et al., 1996; Schulze and Wucherpfennig, 2012). Taken together, the complexity of antigen processing and limited availability of model immunodominant antigens has posed a barrier to understanding the biochemical features of antigenicity.

**[0006]** Historically, defining immunodominant CD4 T cell epitopes has been empirical. More recently, unbiased approaches for epitope discovery have been introduced. In this context, autologous self-epitopes have been identified by immunoaffinity purification of MHCII-associated peptides followed by Edman degradation sequencing or mass spectrometry (Chicz et al., 1993; Chicz et al., 1992; Hunt et al., 1992; Lippolis et al., 2002; Mommen et al., 2016; Rudensky et al., 1991; Sette et al., 1992; Sofiron et al., 2016). Later advancements in mass spectrometry allowed for identification of tumor antigens (Depontieu et al., 2009) and autoantigens (Seamons et al., 2003; Suri et al., 2008). While detection of MHCII-associated peptides derived from exogenous proteins has been achieved (Nelson et al., 1992), identifying epitopes from complex microorganisms remains challenging. Moreover, quantitative assessment of T cell responses to novel epitopes faces technical limitations.

**[0007]** Citation or identification of any document in this application is not an admission that such document is available as prior art to the present invention.

## SUMMARY

**[0008]** It is an objective of the present invention to provide novel methods and computer implemented systems to predict MHCII binding peptides using only a genome sequence of a target cell type or target pathogen (Bacteria originated **T** cell antigen (BOTA)). It is another objective of the present invention to provide novel methods and systems for identifying immunodominant epitopes for any target cell type or target pathogen and for any MHCII allele. In certain example embodiments, Applicants developed methods and systems that utilize a deep neural network that is trained using peptidomic data obtained by isolating MHC II complexes from antigen presenting cells (e.g., dendritic cells). In certain example embodiments, provided is a broadly applicable technology platform for antigen discovery and specification of immunodominance hierarchies.

**[0009]** In one aspect, the present invention provides for a method of preparing one or more peptides for an immunological composition comprising: identifying MHCII antigenic epitopes for a pathogen, commensal microorganism or diseased cell; and formulating an immunological

composition comprising one or more of the identified epitopes, wherein identifying MHCII antigenic epitopes comprises generating, by a processor, a set of candidate antigens from one or more input genome sequences derived from a target pathogen, commensal microorganism or diseased cell, and generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network. The method may further comprise sequencing a biological sample to generate one or more input genome sequences. The deep neural network may be trained using a set of MHCII-presented peptides bound to antigen presenting cells. The antigen presenting cells may be dendritic cells.

**[0010]** In certain embodiments, generating a set of candidate antigens from one or more input genome sequences comprises: predicting, by the processor, genes from an input genome sequence; and defining, by the processor, a set of candidate antigens from the set of predicted genes based on one or more of protein cellular location, transmembrane structure, and domain distribution. Defining, by the processor, a set of candidate antigens from the set of predicted genes, may comprise: selecting surface and secreted proteins; masking intracellular regions and transmembrane domains of the surface and extracellular proteins; excluding regions that are within domains of less than 30 amino acids; excluding regions between a series of inaccessible domains; and excluding inter-domain regions between a series of adjacent domains wherein the inter-domain regions are less than or equal to 20 amino acids. The method may further comprise masking up to 20 amino acids flanking the intracellular regions and transmembrane domains. In certain embodiments, 8 amino acids flanking the intracellular regions and transmembrane domains are masked. The surface protein may be a cell wall protein. The surface protein may be an outer membrane protein. The antigenic epitopes may be derived from candidate antigens more than 20 amino acids away from a transmembrane domain. The candidate antigens may comprise a tertiary structure required for proteolytic enzyme accessibility. The candidate antigens may comprise a defined MHCII-binding motif in the primary amino acid sequence of the predicted genes. The defined MHCII-binding motif may be the I-$A^b$-binding motif. The antigenic epitopes may comprise a tertiary structure required for accessibility to a MHCII binding groove.

**[0011]** In certain embodiments, generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network comprises: encoding each amino acid of an input candidate antigen into a p-dimensional binary vector; converting the binary vector to a descriptor matrix S; convoluting the descriptor matrix S to a scoring matrix X; and transforming the scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for a neural network.

**[0012]** In certain embodiments, the antigenic epitopes are specific for an HLA type.

**[0013]** In certain embodiments, the method further comprises: isolating MHCII-peptide complexes from antigen presenting cells; identifying bound peptides from the MHCII-peptide complexes; and training the deep neural network using the set of identified peptides. The antigen presenting cells may be exposed to a target pathogen, commensal microorganism or diseased cell. The antigen presenting cells may be dendritic cells.

**[0014]** In certain embodiments, the candidate antigens are expressed in the pathogen, commensal microorganism or diseased cell.

**[0015]** In certain embodiments, the pathogen is selected from the group consisting of a bacterium, a virus, a protozoon, and an allergen.

**[0016]** In certain embodiments, the diseased cell is a cancer cell. The cancer cell may be obtained from a subject. The method may further comprise identifying non-silent tumor specific somatic mutations from the subject specific cancer cell, wherein the ranked set of antigenic epitopes are generated from said mutations.

**[0017]** In certain embodiments, the diseased cell is associated with an autoimmune disease.

**[0018]** In certain embodiments, the pathogen comprises a bacteria belonging to a family selected from the group consisting of Bacillus, Bartonella, Bordetella, Borrelia, Brucella, Campylobacter, Chlamydia and Chlamydophila, Clostridium, Corynebacterium, Enterococcus, Escherichia, Francisella, Haemophilus, Helicobacter, Legionella, Leptospira, Listeria, Mycobacterium, Mycoplasma, Neisseria, Pseudomonas, Rickettsia, Salmonella, Shigella, Staphylococcus, Streptococcus, Treponema, Ureaplasma, Vibrio, and Yersinia.

**[0019]** In certain embodiments, the commensal microorganism comprises a bacterium belonging to a genus selected from the group consisting of *Bacteroides, Clostridium, Faecalibacterium, Eubacterium, Ruminococcus, Peptococcus, Peptostreptococcus, Bifidobacterium, Lactobacillus* and *Akkermansia;* or a fungus selected from the group consisting of *Candida, Saccharomyces, Aspergillus, Penicillium, Rhodotorula, Trametes, Pleospora, Sclerotinia, Bullera,* and *Galactomyces.*

**[0020]** In certain embodiments, the antigen presenting cell comprises a subject specific MHCII allele. The antigen presenting cell may comprise a human MHCII allele.

**[0021]** In certain embodiments, the one or more input genome sequences is obtained by whole genome or whole exome sequencing.

**[0022]** In certain embodiments, the method further comprises detecting whether one or more of the antigenic epitopes is present in a sample from a subject suffering from an infection, autoimmune disease, allergy, or cancer.

**[0023]** In certain embodiments, the antigenic epitopes are about 9 to 20 amino acids in length.

**[0024]** In certain embodiments, the deep neural network is trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 1-14979. The deep neural network may be trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 14980-15027.

**[0025]** In certain embodiments, the immunological composition is a protective vaccine or tolerizing vaccine composition comprising one or more of the antigenic epitopes. The vaccine composition may be directed to a bacterium, a virus or a protozoon. The vaccine composition may be directed to a cancer. The vaccine composition may be directed to an autoimmune disease or allergy, whereby administration of the vaccine induces tolerance to the one or more antigenic epitopes.

**[0026]** In certain embodiments, the vaccine composition is directed to *Listeria*. The vaccine composition may comprise one or more *Listeria* peptides selected from the peptides listed in Table 1. The one or more *Listeria* peptides may be derived from lmo0202, lmo2558, lmo2185, or lmo0135. The one or more *Listeria* peptides may be selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984 (APGQETQHYYGLP VAD SAIDR), SEQ ID NO: 14986 (ADFRYVFDTAK ATAAS SYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT).

**[0027]** In certain embodiments, the vaccine composition comprises autologous dendritic cells or antigen presenting cells pulsed with the one or more epitopes.

**[0028]** In another aspect, the present invention provides for a method of identifying one or more immunodominant epitopes comprising: expressing in a first population of immune cells one or more peptides for an immunological composition prepared according to claim 1, wherein the first population of immune cells comprise a detectable reporter gene; expressing in a second population of CD4+ immune cells a TCRαβ library expressing TCRαβ pairs identified in a subject suffering from an infectious disease, autoimmunity, cancer or allergy; incubating the first population of immune cells with the second population of immune cells; sorting immune cells positive for the detectable reporter gene; and identifying peptides bound to MHCII in immune cells positive for the detectable reporter gene. The TCRαβ pairs may be identified by T cell profiling. The TCRαβ pairs may be determined by targeted single cell RNA-seq (TCRseq). The T cells may be analyzed by RNA-seq to determine single cells having a specific cell state and TCRαβ pairs are identified in the T cells having the specific cell state.

The specific cell state may be a protective phenotype or a pathogenic phenotype. The reporter may be an inducible fluorescent marker protein, wherein the marker protein is induced when a TCR$\alpha\beta$ pair detects an MHCII epitope. The fluorescent marker may be under control of the IL-2 promoter. The method may further comprise formulating a vaccine composition comprising one or more of the immunodominant epitopes. The vaccine may be a protective vaccine or a tolerizing vaccine. The vaccine may comprise autologous dendritic cells or antigen presenting cells pulsed with one or more of the immunodominant epitopes.

**[0029]** In another aspect, the present invention provides for a method of identifying peptide antigens from a live bacterial pathogen comprising: isolating MHCII-peptide complexes from dendritic cells exposed to a bacterial pathogen; isolating the peptides from the MHCII-peptide complexes; and sequencing the isolated peptides.

**[0030]** In another aspect, the present invention provides for a method of determining immune related health status in a subject comprising: preparing one or more peptides for an immunological composition according to any embodiment herein; exposing the one or more peptides to immune cells obtained from the subject; and measuring cytokine secretion. The cytokines measured may comprise IL-2, IFN-$\gamma$, IL-17, and/or IL-10. The method may further comprise measuring immune cell types reactive to the epitopes in the subject. The immune cell types may comprise Treg, Thl, Thl7 and/or Th2 cells. The immune cells may be measured using MHCII tetramers.

**[0031]** In another aspect, the present invention provides for a method of determining immune related health status in a subject comprising: preparing one or more peptides for an immunological composition according any embodiment herein; and measuring immune cell types reactive to the epitopes in the subject. The immune cell types may comprise Treg, Thl, Thl7 and/or Th2 cells. The immune cells may be measured using MHCII tetramers.

**[0032]** In another aspect, the present invention provides for a method of constructing a deep neural network for identifying MHCII antigenic epitopes comprising: isolating MHCII-peptide complexes from antigen presenting cells expressing a MHCII type; isolating peptides from the MHCII-peptide complexes; sequencing the isolated peptides; and training a deep neural network for predicting antigenic epitopes for the MHCII type using the set of isolated peptides. The deep neural network may be trained by testing randomly initialized parameters for the peptides. The parameters may be tested in a three-fold cross validation scheme. The parameters may comprise cellular localization, inter-domain structure, domain size and/or

tertiary structure. The cellular localization may comprise peptides derived from extracellular, intracellular or transmembrane proteins.

**[0033]** In another aspect, the present invention provides for a method for identifying MHCII antigenic epitopes comprising: generating, by a processor, a set of candidate antigens from one or more input genome sequences; and generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network, wherein the deep neural network is trained using a set of MHCII-presented peptides isolated from antigen presenting cells. The antigen presenting cells may be dendritic cells.

**[0034]** In certain example embodiments, generating, by the processor, a set of candidate antigens from one or more input genome sequences may comprise: predicting, by the processor, genes from an input genome sequence; and defining, by the processor, a set of candidate antigens from the set of predicted genes based on one or more of protein cellular location, transmembrane structure, and domain distribution.

**[0035]** In certain example embodiments, defining, by the processor, a set of candidate antigens from the set of predicted genes, may comprise: selecting surface and secreted proteins; masking intracellular regions and transmembrane domains of the surface proteins; excluding regions that are within domains of less than 30 amino acids; excluding inter-domain regions between transmembrane domains or domains of less than 30 amino acids; and/or excluding inter-domain regions between a series of adjacent domains wherein the inter-domain regions are less than or equal to 20 amino acids. The method may further comprise masking up to 20 amino acids flanking the intracellular regions and transmembrane domains. In certain embodiments, 8 amino acids flanking the intracellular regions and transmembrane domains are masked. In other embodiments, regions located in or between inaccessible domains are excluded. The term "inaccessible domains" as used herein refers to domains that are (1) intracellular, (2) within a transmembrane domain plus 8 amino acids flanking each end, or (3) any domain identified by PFAM that is less than 30 amino acids in length. These criteria were informed by the training dataset (endogenous mouse peptides bound to MHCII). The metric used for domain mapping may be based on other features, such as the distance to the upstream and downstream domains, the density of flanking domains and the size of the domain. MHCII epitopes may be present between small domains, but not present in small domains.

**[0036]** The surface protein may be a cell wall protein. The surface protein may be an outer membrane protein. The antigenic epitopes may be derived from candidate antigens more than 20 amino acids away from a transmembrane domain. The candidate antigens may comprise a tertiary structure required for proteolytic enzyme accessibility (e.g., solvent exposed epitopes

rather than buried epitopes). In certain embodiments, protein regions predicted to be accessible by proteases are prioritized. The candidate antigens may comprise a defined MHCII-binding motif in the primary amino acid sequence of the predicted genes. This parameter may be used before, during or after the candidate antigens are provided to the neural network. The defined MHCII-binding motif may be the I-$A^b$-binding motif. The antigenic epitopes may comprise a tertiary structure required for accessibility to a MHCII binding groove.

**[0037]** Generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network may comprise: encoding each amino acid of an input candidate antigen into a p-dimensional binary vector; converting the binary vector to a descriptor matrix S; convoluting the descriptor matrix S to a scoring matrix X; and transforming the scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for a neural network.

**[0038]** The method of any embodiment described herein may further comprise: isolating MHCII-peptide complexes from antigen presenting cells; isolating the peptides from the MHCII-peptide complexes; sequencing the isolated peptides; and training the deep neural network using the set of isolated peptides. The antigen presenting cells may be exposed to a target cell type or target pathogen. The antigen presenting cells may be dendritic cells.

**[0039]** The input genome sequence may be derived from a target cell type or target pathogen. The candidate antigens may be expressed in the target cell type or target pathogen. The target cell type or target pathogen of any embodiment herein may be selected from the group consisting of a bacterium, a virus, a protozoa, an allergen and a diseased cell.

**[0040]** The diseased cell may be a cancer cell. The cancer cell may be obtained from a subject in need thereof. The method may further comprise identifying non-silent tumor specific somatic mutations from the subject specific cancer cell, wherein the ranked set of antigenic epitopes are generated from said mutations.

**[0041]** The diseased cell may be associated with an autoimmune disease.

**[0042]** The target pathogen may comprise a bacteria belonging to a family selected from the group consisting of Bacillus, Bartonella, Bordetella, Borrelia, Brucella, Campylobacter, Chlamydia and Chlamydophila, Clostridium, Corynebacterium, Enterococcus, Escherichia, Francisella, Haemophilus, Helicobacter, Legionella, Leptospira, Listeria, Mycobacterium, Mycoplasma, Neisseria, Pseudomonas, Rickettsia, Salmonella, Shigella, Staphylococcus, Streptococcus, Treponema, Ureaplasma, Vibrio, and Yersinia.

**[0043]** The antigen presenting cell may comprise a subject specific MHCII allele. The antigen presenting cell may comprise a human MHCII allele. The one or more input genome

sequences may be obtained by whole genome or whole exome sequencing. The method may further comprise detecting whether one or more of the antigenic epitopes is present in a sample from a subject suffering from an infection, autoimmune disease, allergy, or cancer. In certain embodiments, such as when identifying immunodominant epitopes for personalized medicine, epitopes that are expressed in a subject in need thereof are selected for an immunogenic composition.

**[0044]** The deep neural network may be trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 1-14979. The deep neural network may be trained using a set of MHCII-presented peptides comprising one or more of SEQ ID Nos: 14980-15027.

**[0045]** In certain embodiments, the deep neural network trained using a set of MHCII-presented peptides isolated from a dendritic cell can be used to predict antigenic epitopes from any input genome sequence, such as from a pathogen or a different organism. Not being bound by a theory, training the neural network with MHCII-presented peptides provides the neural network with antigenic determinants for generally predicting MHCII binders. The present invention thus provides novel determinants of antigenicity previously not known.

**[0046]** The method may further comprise preparing a vaccine composition comprising one or more antigenic epitopes from the set of antigenic epitopes generated by the processor. The antigenic epitopes may be about 9 to 20 amino acids in length. The vaccine composition may be directed to a bacterium, a virus or a protozoa. The vaccine composition may be directed to a cancer. The vaccine composition may be directed to an autoimmune disease, whereby administration of the vaccine induces tolerance to the antigenic epitope. The vaccine composition may be directed to *Listeria.* The vaccine composition may comprise one or more *Listeria* peptides selected from the peptides listed in Table 1. The one or more *Listeria* peptides may be derived from lmo0202, lmo2558, lmo2185, or lmo0135. The one or more *Listeria* peptides may be selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984 (APGQETQHYYGLP VAD SAIDR), SEQ ID NO: 14986 (ADFRYVFDT AKATAA SSYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT). Not being bound by a theory, the vaccine may be directed to (1) any microbial pathogen, (2) any commensal microorganism, (3) any food allergen, (4) any tumor/malignancy, (5) any host tissue that is targeted by adaptive immune system in autoimmune disease. The method may be used to prepare tolerizing vaccines or protective vaccines. The vaccine may comprise autologous dendritic ceils or antigen presenting ceils pulsed with the one or more peptides.

**[0047]** In another aspect, the present invention provides for a method of identifying immunodominant epitopes comprising: expressing in a first population of immune cells a peptide MHCII library comprising antigenic epitopes identified according to the method of any embodiment herein, wherein the first population of immune cells comprise a detectable reporter gene; expressing in a second population of CD4+ immune cells a TCRab library expressing TCRab pairs identified in a subject suffering from an infectious disease, autoimmunity, cancer or allergy; incubating the first population of immune cells with the second population of immune cells; sorting immune cells positive for the detectable reporter gene; and identifying peptides bound to MHCII in immune cells positive for the detectable reporter gene. The TCRab pairs may be identified by single cell profiling. The TCRab pairs may be determined by targeted single cell RNA-seq (TCRseq). The T cells may be analyzed by RNA-seq to determine single cells that are activated and TCRab pairs are identified in the activated T cells. The reporter may be an inducible fluorescent marker protein, wherein the marker protein is induced when a TCRab pair detects a MHCII epitope. The fluorescent marker may be under control of the IL-2 promoter. The method may further comprise administering to the subject a vaccine comprising one or more of the immunodominant epitopes. The vaccine may be a protective vaccine or a tolerizing vaccine. The vaccine may comprise autologous dendritic cells or antigen presenting cells pulsed with one or more of the immunodominant epitopes.

**[0048]** In another aspect, the present invention provides for a method of identifying peptide antigens from a live bacterial pathogen comprising: isolating MHCII-peptide complexes from dendritic cells exposed to a bacterial pathogen; isolating the peptides from the MHCII-peptide complexes; and sequencing the isolated peptides.

**[0049]** In another aspect, the present invention provides for a peptide set for training a MHCII neural network comprising one or more of SEQ ID NOs: 1-14979.

**[0050]** In another aspect, the present invention provides for a peptide set for training a MHCII neural network comprising one or more of SEQ ID NOs: 14980-15027.

**[0051]** In another aspect, the present invention provides for a *Listeria* vaccine comprising one or more peptides selected from the peptides listed in Table 1. The one or more peptides may be derived from lmo0202, lmo2558, lmo2185, or lmo0135. The one or more *Listeria* peptides may be selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984 (APGQETQHYYGLP VAD SAIDR), SEQ ID NO: 14986 (ADFRYVFDT AKATAA SSYPG), and SEQ ID NO: 14997

(VDDTTVKFTLPTVAPAFENT). The vaccine may comprise autologous dendritic cells or antigen presenting cells pulsed with the one or more peptides.

**[0052]** It is noted that in this disclosure and particularly in the claims and/or paragraphs, terms such as "comprises", "comprised", "comprising" and the like can have the meaning attributed to it in U.S. Patent law; e.g., they can mean "includes", "included", "including", and the like; and that terms such as "consisting essentially of and "consists essentially of have the meaning ascribed to them in U.S. Patent law, e.g., they allow for elements not explicitly recited, but exclude elements that are found in the prior art or that affect a basic or novel characteristic of the invention.

**[0053]** These and other aspects, objects, features, and advantages of the example embodiments will become apparent to those having ordinary skill in the art upon consideration of the following detailed description of illustrated example embodiments.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0054]** **FIG. 1** - block diagram depicting a system for identifying MHCII antigenic epitopes.

**[0055]** **FIG. 2** - block diagram depicting a system for a deep neural network for predicting binding of peptides to MHCII molecules.

**[0056]** **FIG. 3** - block diagram depicting a method for identifying MHCII antigenic epitopes.

**[0057]** **FIG. 4** - block diagram depicting a method for generating a set of candidate antigens from one or more input genome sequences.

**[0058]** **FIG. 5** - block diagram depicting a method for generating a ranked set of antigenic epitopes from the candidate antigens using a deep neural network.

**[0059]** **FIG. 6** - block diagram depicting a computing machine and a module, in accordance with certain example embodiments.

**[0060]** **FIG. 7 - MHCII peptidomics in primary murine dendritic cells results in more than 3,700 distinct peptide identifications and defines the I-A$^{b}$-binding motif. Fig. 7A** is an example experimental workflow for immunopurification and sequencing of MHCII-associated peptides from murine dendritic cells. MHCII-peptide complexes were immunopurified from WT and *Atgl6H*$^{\sim\sim}$ cells. Associated peptides were then acid-eluted, labeled with iTRAQ4 reagents, desalted with SCX and C18, and analyzed using high resolution LC-MS/MS. **Fig. 7B** is an example database search strategy for MHCII-peptide sequencing. All MS/MS spectra were searched against a database containing mouse proteins using

Spectrum Mill software with a "no enzyme" specificity. Mouse peptides were validated using a 1% FDR cutoff, and the total numbers of peptides quantified across all samples were reported. **Fig. 7C** is the I-Ab-binding motif was derived from endogenous mouse peptides bound to MHCII. Heatmap shading represents the frequencies of each amino acid at each respective position.

**[0061] FIG. 8 - Antigen processing pathways and epitope features revealed by MHCII peptidomics. Fig. 8A** shows that deficiency in the autophagy protein Atgl611 dramatically skews the spectrum of MHCII-associated peptides. MHCII-bound peptides quantified in *Atg6H*$^{-/-}$ dendritic cells relative to wild type (WT). Replicate (rep) samples were compared based on log2 fold change (FC) between mouse strains. Each dot represents a unique peptide sequence. Peptides that were observed to be significantly upregulated or downregulated are shown, while peptide measurements that were not reproducible across both biological replicates are also shown. Dot plot axes: Log2FC. Histogram axes: number of distinct peptides. **Fig. 8B** shows the abundance and subcellular sources of MHCII-associated peptides derived from *Atgl6H*$^{-/-}$ and WT dendritic cells. **Fig. 8C** shows epitope mapping relative to domain structure of endogenous antigens indicates preferential presentation of epitopes derived from the lumenal/extracellular domains of transmembrane proteins and epitopes positioned between structurally-defined domains. **Fig. 8D** shows immunodominant epitope prediction with BOTA. Workflow of BOTA with input as genome and output as a binding score. The upper panel shows the extraction of candidate peptides (SEQ ID NO: 15045-15048), and the lower panel shows the deep neural network core of the BOTA algorithm to assign a binding score to each candidate peptide. **Fig. 8E** shows MHCII-associated peptides derived from *Atgl6H*$^{-/-}$ and WT dendritic cells. Shown are SEQ ID NO: 637, 3741, 3753, 3756, 7518, 9956, 1231 1, 13883, and 15049-15100, respectively.

**[0062] FIG. 9 - Validation of BOTA epitope predictions with MHCII peptidomics. Fig. 9A** shows an experimental workflow for immunopurification and sequencing of MHCII-associated peptides from murine dendritic cells. MHCII-peptide complexes were immunopurified from WT cells after a 10-minute or 6-hour *Listeria* treatment. Associated peptides were then acid-eluted, labeled with iTRAQ4 reagents, desalted with SCX and C18, and analyzed using high resolution LC-MS/MS. **Fig. 9B** shows MHCII-bound peptides detected before and after *Listeria* exposure. Biological replicates (rep) were compared based on log2 fold change (FC) between time 10 min and 6 hr after exposure to bacteria. Each dot represents a unique peptide sequence. Peptides that were observed to be significantly upregulated or downregulated are shown, while peptide measurements that were not

reproducible across both biological replicates are also shown. Dot plot axes: Log2FC. Histogram axes: number of distinct peptides. **Fig. 9C** shows predictions for Listeria epitopes were made using the deep neural network core of BOTA. **Fig. 9D** shows that the BOTA model pre-training accuracy plateaus after 200 epochs in cross-validation. The model was trained using the mouse peptides captured in BMDCs infected by Listeria (line); in contrast, the same model trained solely on Immune Epitope Database (IEDB) data reached a plateau at approximately 70%, signifying a 15% gap in accuracy. **Fig. 9E** shows a comparison of predictions for *Listeria* epitopes in proteins identified by proteomics. Peptides are split into categories based on the protein's subcellular localization using PSORTb.

**[0063] FIG. 10 - BOTA and MHCII peptidomics accurately predict immunodominance *in vivo.*** **Fig. 10A** shows epitope mapping and domain structure of *Listeria* antigens indicate preferential presentation of surface-exposed and secreted proteins. **Fig. 10B** shows that immunodominance of epitopes was determined by infecting mice with *Listeria.* At day 7, splenocytes were harvested and restimulated with the indicated peptides for quantification of the T cell response by IFNy ELISPOT. Data represent the mean number of spots per $1x10^5$ CD4 T cells ± sd for n = 6 mice. **Fig. 10 C, D and E** show that immunodominance *in vivo* correlates with fold change (FC) of *Listeria* peptides quantified by MHCII peptidomics, and to a lesser extent, with mRNA expression (normalized microarray probe intensity) of the corresponding peptide-encoding genes.

**[0064] FIG. 11 - Single cell TCR sequencing defines the clonal architecture of the *anti-Listeria* response.** Mice were inoculated with Listeria by i.p. injection. Six days later, CD4+CD25+CD69+ T cells were FACS-sorted from spleens for single cell RNAseq and TCRseq. **Fig. 11A** shows a tSNE plot derived from T cell transcriptomes identifies 3 distinct cell states that cluster according to unique signatures. Each dot represents a single cell that is color coded according to gene signature. **Fig. 11B** shows that TCRseq defines the CD4 T cell response to Listeria. Each dot represents a single cell that is shaded according to expression of Ifng. **Fig. llC** shows that each dot represents a single cell that is shaded according to expression of Tbx21. **Fig. 11D** shows circos plots of the linkages between the two segments in the alpha chain (top) and beta chain (bottom). Segment J's coded in grey and segment V's are differentially shaded. Ribbons link two chains with thickness proportional to the number of corresponding V/J pairs observed, and with the shading identical to the V chain.

**[0065] FIG. 12 - Single cell RNAseq integrates T cell phenotype with TCR repertoire in the *Listeria* response.** Mice were inoculated with Listeria by i.p. injection on days 0 and 11. On day 18, $FSC^{HI}CD4^{+}CD8^{-}B220^{-}$MHCIT T cells were FACS-sorted from spleens for single

cell RNAseq and TCRseq. Sorted T cells were derived from 2 mice, sequenced separately, and combined for analysis. **Fig. 12A** shows a tSNE plot derived from T cell transcriptomes identifies distinct cell states that cluster according to unique signatures. Each dot represents a single cell that is shaded according to gene signature. **Fig. 12B** shows violin plots displaying Teff signature score derived from ImmGen. Each dot represents a single cell classified by clusters defined by tSNE. **Fig. 12C** shows circos plots of the linkages between the TCR alpha chain CDR3 and TCR beta chain CDR3. Ribbons link two chains with thickness proportional to the number of corresponding TCR pairs observed. Dominant TCR clones are labeled according to TCR gene segment usage. Starting with the sequence listed next to the TRBV5/TRBJ2-7 TCR, the plot lists SEQ ID NO:15101-15197, going in clock-wise order.

**[0066] FIG. 13 - Specifying immunodominance by screening TCRs for reactivity with Listeria epitopes predicted by BOTA. Fig. 13A** shows the screen overview. BW5147_B7-4 cells were transduced to express peptide epitopes fused in-frame with the I-A$^b$beta chain bearing CD3zeta cytoplasmic domains. BW5147_CD4-28 cells were transduced to express chimeric single chain TCRs bearing the transmembrane and cytoplasmic domains of CD3zeta. In this coculture system, cognate antigen recognition results in T cell activation characterized by IL-2 production. **Fig. 13B** shows that TCRs associated with high Ifng expression, as identified in CD4 T cells from Listeria-infected mice, were screened against Listeria epitopes. The positive control OT2 TCR reacted robustly with OT2 peptide I-A$^b$. T cell activation was measured in duplicate by induction of Nur77 expression relative to beta actin. The right side of panel B lists SEQ ID NO:15198-15200, from left to right.

**[0067] FIG. 14 - Specifying antigen-reactivity by screening TCRs for reactivity with *Listeria* epitopes predicted by BOTA. Fig. 14A** shows the screen overview. HEK 293T cells were transfected to express peptide epitopes fused in-frame with the I-A$^b\beta$ chain bearing CD3$\zeta$ cytoplasmic domains. BW5147_CD4-28 cells were transduced to express chimeric single chain TCRs bearing the transmembrane and cytoplasmic domains of CD3$\zeta$. In this coculture system, cognate antigen recognition results in T cell activation characterized by production of IL-2. **Fig. 14B** shows that the most abundant TCR identified in mice infected with Listeria (lmo_R6) was screened for reactivity against Listeria epitopes as described above. IL-2 was detected in culture supernatant by cytometric bead array. As controls, OT2 TCR (reactive with Ova) and LLO_l18 TCR (reactive with LLO) were included. Shown are SEQ ID NO:15200-15204, from left to right. **Fig. 14C** shows that HEK 293T cells were transfected with constructs encoding single chain TCRs. Cells were analyzed by FACS for expression of TCRb and binding to LLO I-A$^b$ tetramers (NEKYAQAYPNVS I-A$^b$) (SEQ ID NO: 15205).

**[0068] FIG. 15 - Computational prediction and validation of a dominant commensal antigen. Fig. 15A** shows that mice were administered DSS to induce colitis prior to analysis by SICC-seq. At day 14, serum was collected and incubated with stool to allow binding of IgG with commensals. IgG-positive and -negative fractions were separated with magnetic beads covalently attached to protein *A*/*G*. The immunogenicity of *Bacteroidales* was demonstrated by IgG-reactivity score (relative abundance in IgG-positive minus IgG-negative fractions) derived from 16s sequencing. **Fig. 15B** shows that BOTA identified SusC, a highly represented epitope within and a 1 cross *Bacteroidales* genomes, including the murine commensal *Muribaculum intestinale.* Splenocytes from naive mice were harvested and stimulated in vitro with the indicated peptide. Cytokines were measured 24 hours later by cytometric bead array. IL-10 production was selectively induced by SusC peptide, suggesting that host T cells recognize and tolerate *Bacteroidales.* Shown are SEQ ID NO: 15206-15208, starting with the top peptide.

**[0069] FIG. 16 - Features of antigenicity.** Epitopes from MHCII peptidomics experiments were mapped back to the endogenous murine proteins from which they derived. If epitopes localized between PFAM domains, inter-domain sizes were plotted (Top). For epitopes that localized within PFAM domains, the domain sizes were plotted (Bottom). The same procedure was performed for Non-epitopes, which were defined as peptides conforming to the I-$A^b$-binding consensus motif (motif score > $5xl0^{-11}$). Epitopes preferentially derived from inter-domain regions that are greater than 20 amino acids in length or from within domains of more than 30 amino acids. A.A.; amino acids.

**[0070] FIG. 17 - Strategy for TCRseq coupled with whole transcriptome sequencing.** Single T cells are FACS-sorted into individual wells of 384-well PCR plates. First-strand cDNA is primed with polydT oligonucleotide appended with adaptor. Template switching during reverse transcription appends the cDNA with well barcodes (XXXXXX), unique molecular identifiers (NNNNNNN), and adapters. Whole transcriptome amplification (WTA) is then performed with PCR primers complementary to the sequencing adapters. After WTA, individual wells are pooled, DNA is purified, and then divided into three fractions to produce sequencing libraries for TCR-alpha, TCR-beta, and 5' digital gene expression (5'DGE). TCR pre-amplification is performed with PCR primers complementary to the 5' adapter and TCR constant domain. A second round TCR PCR is then performed with PCR primers containing full-length P7 Illumina sequencing adapters that are complementary to the 5' template DNA and a nested primer complementary to the TCR constant domain and appended with P5 illumina sequencing adapters. TCR libraries were sequenced with a 608 cycle MiSeq run.

5'DGE libraries are produced by tagmentation of pooled WTA reactions. A 5' cDNA enrichment PCR adds full length P7 Illumina sequencing adapters and N5 Nextera sequencing adapters. 5'DGE libraries were sequenced with a 75 cycle NextSeq run.

**[0071] FIG. 18 - BOTA Applications:** Strategy for specifying immunodominance hierarchies at scale.

**[0072] FIG. 19 - TCR Finger-Printing. Fig. 19A** shows an example schematic overview. Peptide-MHCn¾ cDNA libraries encoding mutated ovalbumin peptides (Ova 323-339) were delivered to BW_B7/4 T cells by lentiviral transduction. These responder cells were cocultured with BW_4-28 stimulator cells expressing a single chain OT2 TCRζ construct. Upon engagement of the OT2 TCR with cognate Ova(323-339) presented by MHOIζζ, responder cells express the activation marker 41bb. Activated MHCII+41bb+ cells were FACS-sorted, and sequencing libraries were constructed by PCR amplification of the Ova-encoding region of lentiviral integrants. **Fig. 19B** shows that Illumina sequencing reads were translated in silico and quantified to compare the abundances of Ova-encoding mutants within the 41bb+ samples relative to unsorted samples. Shown are SEQ ID NO: 15209-15214, from left to right. **Fig. 19C** shows that a glutamate at amino acid position 11 in Ova(323-339) is strictly required for functional interaction with the OT2 TCR. This position represents an outward-facing TCR contact reside within the core MHCII-binding register of Ova(323-339). In contrast, position 2 of Ova(323-339) falls within a region flanking the core MHCII-binding motif and can accommodate several different amino acids while retaining functional interaction with the OT2 TCR. This observation is consistent with the notion that position 2 does not contribute to recognition by the OT2 TCR. Shown are SEQ ID NO: 15215-15218, starting with the top sequence.

**[0073] FIG. 20 -** Illustrates immune cell type reactivity and health status.

**[0074] FIG. 21 -** Illustrates a clinical immunology pipeline for identifying HLA specific epitopes for use in measuring commensal-reactivity/tolerance profiles and for tracking commensal T cells in patients using the epitopes.

**[0075] FIG. 22 -** Illustrates features of immunogenicity for SusC epitopes associated with Treg reactivity in healthy mice. Shown are SEQ ID NO: 15219-15234, starting with the top sequence.

## DETAILED DESCRIPTION OF THE EXAMPLE EMBODIMENTS

### General Definitions

**[0076]** Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure pertains. Definitions of common terms and techniques in molecular biology may be found in Molecular Cloning: A Laboratory Manual, 2nd edition (1989) (Sambrook, Fritsch, and Maniatis); Molecular Cloning: A Laboratory Manual, 4th edition (2012) (Green and Sambrook); Current Protocols in Molecular Biology (1987) (F.M. Ausubel et al. eds.); the series Methods in Enzymology (Academic Press, Inc.): PCR 2: A Practical Approach (1995) (M.J. MacPherson, B.D. Hames, and G.R. Taylor eds.): Antibodies, A Laboraotry Manual (1988) (Harlow and Lane, eds.): Antibodies A Laboraotry Manual, 2nd edition 2013 (E.A. Greenfield ed.); Animal Cell Culture (1987) (R.I. Freshney, ed.); Benjamin Lewin, Genes IX, published by Jones and Bartlet, 2008 (ISBN 0763752223); Kendrew *et al.* (eds.), The Encyclopedia of Molecular Biology, published by Blackwell Science Ltd., 1994 (ISBN 0632021829); Robert A. Meyers (ed.), Molecular Biology and Biotechnology: a Comprehensive Desk Reference, published by VCH Publishers, Inc., 1995 (ISBN 9780471 185710); Singleton *etal.,* Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, N.Y. 1994), March, Advanced Organic Chemistry Reactions, Mechanisms and Structure 4th ed., John Wiley & Sons (New York, N.Y. 1992); and Marten H. Hofker and Jan van Deursen, Transgenic Mouse Methods and Protocols, 2nd edition (201 1) .

**[0077]** As used herein, the singular forms "a", "an", and "the" include both singular and plural referents unless the context clearly dictates otherwise.

**[0078]** The term "optional" or "optionally" means that the subsequent described event, circumstance or substituent may or may not occur, and that the description includes instances where the event or circumstance occurs and instances where it does not.

**[0079]** The recitation of numerical ranges by endpoints includes all numbers and fractions subsumed within the respective ranges, as well as the recited endpoints.

**[0080]** The terms "about" or "approximately" as used herein when referring to a measurable value such as a parameter, an amount, a temporal duration, and the like, are meant to encompass variations of and from the specified value, such as variations of +/-10% or less, *+1-5%* or less, +/-1% or less, and +/-0.1% or less of and from the specified value, insofar such variations are appropriate to perform in the disclosed invention. It is to be understood that the value to which the modifier "about" or "approximately" refers is itself also specifically, and preferably, disclosed.

**[0081]** Reference throughout this specification to "one embodiment", "an embodiment," "an example embodiment," means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in one embodiment," "in an embodiment," or "an example embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to a person skilled in the art from this disclosure, in one or more embodiments. Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention. For example, in the appended claims, any of the claimed embodiments can be used in any combination.

**[0082]** All publications, published patent documents, and patent applications cited in this application are indicative of the level of skill in the art(s) to which the application pertains. All publications, published patent documents, and patent applications cited herein are hereby incorporated by reference to the same extent as though each individual publication, published patent document, or patent application was specifically and individually indicated as being incorporated by reference.

**Overview**

**[0083]** Defining the immunodominance hierarchy of T cell epitopes remains a significant challenge in the context of infectious disease, commensal immunity, autoimmunity, and immune oncology. Even for some of the most widespread bacterial pathogens, little is known about which antigens drive protective CD4 T cell responses. In one aspect, embodiments herein provide computer-implemented techniques for predicting MHCII binding epitopes using sequencing data. In certain example embodiments, the sequencing data is whole genome data, whole exome sequencing data (WES), RNA-seq data, targeted exome sequencing data, or any form of sequencing data that allows gene prediction at either the exome, genome, or RNA levels. In certain example embodiments, it is not necessary to have a complete genome to identify all useful epitopes. For ease of reference, the sequencing data as described above may be used interchangeably. In certain embodiments, only 90% of a genome or 90% of an exome is required to predict MHCII binding epitopes. In certain embodiments, predictions can be made from metagenomics (i.e., genomic sequencing from a complex mixture of microbes in stool) in which coverage of any given microbe's genome may be as low as 1%. In certain embodiments, de novo predictions can be made from sequencing reads directly from

metagenomics or transcriptomics without mapping to a genome (e.g., essentially 0% known genome).

**[0084]** As a consequence of deep profiling of the MHC II immunopeptidome, a rich dataset was generated for identifying key features of antigenicity and derivation of the computer-implemented methods disclosed herein. Accordingly, the computer-implemented methods disclosed herein incorporate several important attributes of immunodominate epitopes revealed by proteomics.

**[0085]** The improvement on predication accuracy of the computer-implemented methods disclosed herein also signifies the successful application of deep neural network solving to a complex biomedical problem. Previous efforts using traditional networks or hidden Markov Models were limited by their ability to extract highly abstract features, thus leading to insufficient insights into epitope prediction. The computer-implemented methods disclosed herein can be used to predict CD4 T cell epitopes for virtually any MHC II allele and any antigen source, including commensal microbes, pathogens, autoantigens, and tumor antigens. The embodiments disclosed herein demonstrate the utility of MHCII peptidomics for training a deep neural network that specifically identifies epitopes with key features associated with immunodominance. Furthermore, the embodiments disclosed herein serve as a powerful tool for unbiased discovery of complex pathogen and for interrogation of host pathways underlying human disease.

**[0086]** In certain embodiments, MHCII binding epitopes may be predicted using known MHCII binding motifs. The known binding motifs may be incorporated into the described systems and methods. Candidate antigens with a known binding motif may be used as input to the deep neural network system 116. In one embodiment, human HLA class II binding motifs may be used (see e.g., Andreatta and Nielsen, "Characterizing the binding motifs of 11 common human HLA-DP and HLA-DQ molecules using NNAlign" Immunology. 2012 Jul; 136(3): 306-31 1).

**[0087]** Turning now to Figures 1-6, in which like numerals represent like (but not necessarily identical) elements throughout the figures, example embodiments are described in detail.

**Example Epitope Prediction System Architectures**

**[0088]** Figure 1 is a block diagram depicting an epitope prediction system 100 for predicting MHCII binding antigens using nucleic acid sequencing data (e.g., BOTA). As depicted in Figure 1, the operating environment 100 includes network devices 105 and 110 that are configured to communicate with one another via one or more networks 120. In some

embodiments, the genomic sequencing device 105 may provide sequencing data directly from real time sequencing reads. In other example embodiments, the genomic sequencing device comprises a storage device 106. The storage device 106 may comprise a database or be linked to a database comprising sequencing files 107. In some embodiments, a user associated with a device must install an application and/or make a feature selection to obtain the benefits of the techniques described herein. The epitope prediction system 110 receives genomic data and outputs a set of identified ranked epitopes from the sequencing data. BOTA scoring reflects the likelihood of a given epitope being immunodominant.

**[0089]** In certain embodiments, sequencing comprises high-throughput (formerly "next-generation") technologies to generate sequencing reads. In DNA sequencing, a read is an inferred sequence of base pairs (or base pair probabilities) corresponding to all or part of a single DNA fragment. A typical sequencing experiment involves fragmentation of the genome into millions of molecules or generating complementary DNA (cDNA) fragments, which are size-selected and ligated to adapters. The set of fragments is referred to as a sequencing library, which is sequenced to produce a set of reads. Methods for constructing sequencing libraries are known in the art (see, e.g., Head et al., Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 2014; 56(2): 61-77). In certain embodiments, the library members (e.g., genomic DNA, cDNA) may include sequencing adaptors that are compatible with use in, e.g., Alumina's reversible terminator method, Roche's pyrosequencing method (454), Life Technologies' sequencing by ligation (the SOLID platform) or Life Technologies' Ion Torrent platform. Examples of such methods are described in the following references: Margulies et al (Nature 2005 437: 376-80); Ronaglii et al (Analytical Biochemistry 1996 242: 84-9); Shendure et al (Science 2005 309: 1728-32); Imelfort et al (Brief Bioinform. 2009 10:609-18); Fox et al (Methods Mol. Biol. 2009; 553:79-108); Appleby et al (Methods Mol. Biol. 2009; 513: 19-39); and Morozova et al (Genomics. 2008 92:255-64), which are incorporated by reference for the general descriptions of the methods and the particular steps of the methods, including all starting products, reagents, and final products for each of the steps.

**[0090]** In certain embodiments, the present invention includes whole genome sequencing. Whole genome sequencing (also known as WGS, full genome sequencing, complete genome sequencing, or entire genome sequencing) is the process of determining the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast.

**[0091]** In certain embodiments, the present invention includes whole exome sequencing. Exome sequencing, also known as whole exome sequencing (WES), is a genomic technique for sequencing all of the protein-coding genes in a genome (known as the exome) (see, e.g., Ng et al., 2009, Nature volume 461, pages 272-276). It consists of two steps: the first step is to select only the subset of DNA that encodes proteins. These regions are known as exons - humans have about 180,000 exons, constituting about 1% of the human genome, or approximately 30 million base pairs. The second step is to sequence the exonic DNA using any high-throughput DNA sequencing technology.

**[0092]** In certain embodiments, targeted sequencing is used in the present invention (see, e.g., Mantere et al., PLoS Genet 12 el005816 2016; and Carneiro et al. BMC Genomics, 2012 13:375). Targeted gene sequencing panels are useful tools for analyzing specific mutations in a given sample. Focused panels contain a select set of genes or gene regions that have known or suspected associations with the disease or phenotype under study.

**[0093]** Each network device 105 and 110 includes a device having a communication module capable of transmitting and receiving data over the network 120. For example, each network device 105 and 110 can include a server, desktop computer, laptop computer, tablet computer, a television with one or more processors embedded therein and / or coupled thereto, smart phone, handheld computer, personal digital assistant ("PDA"), or any other wired or wireless, processor-driven device.

**[0094]** The genomic sequencing device 105 may generate sequence data files 107 comprising information on the coding regions of genes within a given biological sample. In one example embodiment, the sequencing device may directly communicate the data file to the epitope prediction system 110 across the network 120 and the epitope prediction and ranking is conducted in line with the sequencing analysis. In another example embodiment, the sequencing data file may be stored on a data storage medium and later uploaded to the epitope prediction system 110 for further analysis.

**[0095]** The epitope prediction system 110 may comprise a peptide prediction module 111, a epitope ranking module 113, a candidate peptide library index 112, and a candidate epitope module 114. The peptide prediction module 111 predicts protein coding genes from the input genomic sequencing data file 107 and generates a list of candidate peptides based on various selection criteria described below. In certain example embodiments, the peptide prediction module 111 stores the candidate peptides in the candidate peptide library index 112.

**[0096]** It will be appreciated that the network connections shown are examples and other means of establishing a communications link between the computers and devices can be used.

Moreover, those having ordinary skill in the art having the benefit of the present disclosure will appreciate that the Epitope Prediction System 110, can have any of several other suitable computer system configurations.

**Example Processes**

**[0097]** The example methods illustrated in FIGS. 2-4 are described hereinafter with respect to the components of the example operating environment 100. The example methods of FIGS. 2-4 may also be performed with other systems and in other environments.

**[0098]** FIG. 2 is a block flow diagram depicting a method 300 for identifying MHCII antigenic epitopes an input genome sequence, in accordance with certain example embodiments.

**[0099]** The input genome sequence according to any embodiment described herein may be derived from any target cell type or target pathogen. In certain embodiments, the input genome sequence may be derived from the target cell type or target pathogen exposed to a dendritic cell in any embodiment described herein. The target cell type or target pathogen may be selected from the group consisting of a bacterium, virus, protozoa, and diseased cell. The diseased cell may be a cancer cell. The cancer cell may be obtained from a subject in need thereof. The method may further comprise identifying non-silent tumor specific somatic mutations from the input genome sequence, wherein the antigenic epitopes are derived from said mutations. Not being bound by a theory, a cancer obtains non-inherited mutations that can produce non-self tumor specific mutations. The tumor specific mutations can produce tumor specific antigens that are ideal for targeting by the immune system. Identifying non-silent tumor specific somatic mutations from the input genome sequence may be determined by sequencing tumor tissue from a subject and comparing the sequence to that of non-tumor tissue from the same subject or from germline tissue from the same subject. The diseased cell may be associated with an autoimmune disease.

**[0100]** In certain example embodiments, the target pathogen is a bacterium. Examples of pathogenic bacteria that can be detected in accordance with the disclosed methods include without limitation any one or more of (or any combination of) *Acinetobacter baumanii, Actinobacillus sp., Actinomycetes, Actinomyces sp.* (such as *Actinomyces israelii* and *Actinomyces naeslundii), Aeromonas sp.* (such as *Aeromonas hydrophila, Aeromonas veronii biovar sobria (Aeromonas sobria),* and *Aeromonas caviae), Anaplasma phagocytophilum, Anaplasma marginale, Alcaligenes xylosoxidans, Acinetobacter baumanii, Actinobacillus actinomycetemcomitans, Bacillus sp.* (such as *Bacillus anthracis, Bacillus cereus, Bacillus subtilis, Bacillus thuringiensis,* and *Bacillus stearothermophilus), Bacteroides sp.* (such as

*Bacteroides fragilis), Bartonella sp.* (such as *Bartonella bacilliformis* and *Bartonella henselae, Bifidobacterium sp., Bordetella sp.* ( such as *Bordetella pertussis, Bordetella parapertussis,* and *Bordetella bronchiseptica), Borrelia sp.* (such as *Borrelia recurrentis,* and *Borrelia burgdorferi), Brucella sp.* (such as *Brucella abortus, Brucella canis, Brucella melintensis* and *Brucella suis), Burkholderia sp.* (such as *Burkholderia pseudomallei* and *Burkholderia cepacia), Campylobacter sp.* (such as *Campylobacter jejuni, Campylobacter coli, Campylobacter lari* and *Campylobacter fetus), Capnocytophaga sp., Cardiobacterium hominis, Chlamydia trachomatis, Chlamydophila pneumoniae, Chlamydophila psittaci, Citrobacter sp. Coxiella burnetii, Corynebacterium sp.* (such as, *Corynebacterium diphtheriae, Corynebacterium jeikeum* and *Corynebacterium), Clostridium sp.* (such as *Clostridium perfringens, Clostridium difficile, Clostridium botulinum* and *Clostridium tetani), Eikenella corrodens, Enterobacter sp.* (such as *Enterobacter aerogenes, Enterobacter agglomerans, Enterobacter cloacae* and *Escherichia coli, including opportunistic Escherichia coli,* such as *enterotoxigenic E. coli, enteroinvasive E. coli, enteropathogenic E. coli, enterohemorrhagic E. coli, enteroaggregative E. coli* and *uropathogenic E. coli) Enterococcus sp.* (such as *Enterococcus faecalis* and *Enterococcus faecium) Ehrlichia sp.* (such as *Ehrlichia chafeensia* and *Ehrlichia canis), Erysipelothrix rhusiopathiae, Eubacterium sp., Francisella tularensis, Fusobacterium nucleatum, Gardnerella vaginalis, Gemella morbillorum, Haemophilus sp.* (such as *Haemophilus influenzae, Haemophilus ducreyi, Haemophilus aegyptius, Haemophilus parainfluenzae, Haemophilus haemolyticus* and *Haemophilus parahaemolyticus, Helicobacter sp.* (such as *Helicobacter pylori, Helicobacter cinaedi* and *Helicobacter fennelliae), Kingella kingii, Klebsiella sp.* ( such as *Klebsiella pneumoniae, Klebsiella granulomatis* and *Klebsiella oxytoca), Lactobacillus sp., Listeria monocytogenes, Leptospira interrogans, Legionella pneumophila, Leptospira interrogans, Peptostreptococcus sp., Mannheimia hemolytica, Moraxella catarrhalis, Morganella sp., Mobiluncus sp., Micrococcus sp. Mycobacterium sp.* (such as *Mycobacterium leprae, Mycobacterium tuberculosis, Mycobacterium paratuberculosis, Mycobacterium intracellulare, Mycobacterium avium, Mycobacterium bovis,* and *Mycobacterium marinum), Mycoplasm sp.* (such as *Mycoplasma pneumoniae, Mycoplasma hominis,* and *Mycoplasma genitalium), Nocardia sp.* (such as *Nocardia asteroides, Nocardia cyriacigeorgica* and *Nocardia brasiliensis), Neisseria sp.* (such as *Neisseria gonorrhoeae* and *Neisseria meningitidis), Pasteurella multocida, Plesiomonas shigelloides. Prevotella sp., Porphyromonas sp., Prevotella melaninogenica, Proteus sp.* (such as *Proteus vulgaris* and *Proteus mirabilis), Providencia sp.* (such as *Providencia alcalifaciens, Providencia rettgeri* and *Providencia stuartii), Pseudomonas aeruginosa, Propionibacterium*

*acnes, Rhodococcus equi, Rickettsia sp.* (such as *Rickettsia rickettsii, Rickettsia akari* and *Rickettsia prowazekii, Orientia tsutsugamushi {formerly: Rickettsia tsutsugamushi)* and *Rickettsia typhi), Rhodococcus sp., Serratia marcescens, Stenotrophomonas maltophilia, Salmonella sp.* (such as *Salmonella enterica, Salmonella typhi, Salmonella paratyphi, Salmonella enteritidis, Salmonella cholerasuis* and *Salmonella typhimurium), Serratia sp.* (such as *Serratia marcesans* and *Serratia liquifaciens), Shigella sp.* (such as *Shigella dysenteriae, Shigella flexneri, Shigella boydii* and *Shigella sonnei), Staphylococcus sp.* (such as *Staphylococcus aureus, Staphylococcus epidermidis, Staphylococcus hemolyticus, Staphylococcus saprophyticus), Streptococcus sp.* (such as *Streptococcus pneumoniae* (for example *chloramphenicol-resistant serotype 4 Streptococcus pneumoniae, spectinomycin-resistant serotype 6B Streptococcus pneumoniae, streptomycin-resistant serotype 9V Streptococcus pneumoniae, erythromycin-resistant serotype 14 Streptococcus pneumoniae, optochin-resistant serotype 14 Streptococcus pneumoniae, rifampicin-resistant serotype 18C Streptococcus pneumoniae, tetracycline-resistant serotype 19F Streptococcus pneumoniae, penicillin-resistant serotype 19F Streptococcus pneumoniae,* and *trimethoprim-resistant serotype 23F Streptococcus pneumoniae, chloramphenicol-resistant serotype 4 Streptococcus pneumoniae, spectinomycin-resistant serotype 6B Streptococcus pneumoniae, streptomycin-resistant serotype 9V Streptococcus pneumoniae, optochin-resistant serotype 14 Streptococcus pneumoniae, rifampicin-resistant serotype 18C Streptococcus pneumoniae, penicillin-resistant serotype 19F Streptococcus pneumoniae, or trimethoprim-resistant serotype 23F Streptococcus pneumoniae), Streptococcus agalactiae, Streptococcus mutans, Streptococcus pyogenes, Group A streptococci, Streptococcus pyogenes, Group B streptococci, Streptococcus agalactiae, Group C streptococci, Streptococcus anginosus, Streptococcus equismilis, Group D streptococci, Streptococcus bovis, Group F streptococci,* and *Streptococcus anginosus Group G streptococci), Spirillum minus, Streptobacillus moniliformi, Treponema sp.* (such as *Treponema carateum, Treponema petenue, Treponema pallidum* and *Treponema endemicum, Tropheryma whippelii, Ureaplasma urealyticum, Veillonella sp., Vibrio sp.* (such as *Vibrio cholerae, Vibrio parahemolyticus, Vibrio vulnificus, Vibrio parahaemolyticus, Vibrio vulnificus, Vibrio alginolyticus, Vibrio mimicus, Vibrio hollisae, Vibrio fluvialis, Vibrio metchnikovii, Vibrio damsela* and *Vibriofurnisii), Yersinia sp.* ( such as *Yersinia enterocolitica, Yersinia pestis,* and *Yersinia pseudotuberculosis)* and *Xanthomonas maltophilia* among others.

**[0101]** In certain example embodiments, the target pathogen is a fungus. Examples of fungi that can be detected in accordance with the disclosed methods include without limitation any one or more of (or any combination of), Aspergillus, Blastomyces, Candidiasis,

Coccidiodomycosis, Cryptococcus neoformans, Cryptococcus gatti, Histoplasma, Mucroymcosis, Pneumocystis, Sporothrix, fungal eye infections ringwork, Exserohilum, and Cladosporium.

**[0102]** In certain example embodiments, the fungus is a yeast. Examples of yeast that can be detected in accordance with disclosed methods include without limitation one or more of (or any combination of), *Aspergillus* species, a *Geotrichum* species, a *Saccharomyces* species, a *Hansenula* species, a *Candida* species, a *Kluyveromyces* species, a *Debaryomyces* species, a *Pichia* species, or combination thereof. In certain example embodiments, the fungus is a mold. Example molds include, but are not limited to, a *Penicillium* species, a *Cladosporium* species, a *Byssochlamys* species, or a combination thereof.

**[0103]** In certain example embodiments, the target pathogen may be a virus. The virus may be a DNA virus, a RNA virus, or a retrovirus. Example of RNA viruses that may be detected include one or more of (or any combination of) Coronaviridae virus, a Picornaviridae virus, a Caliciviridae virus, a Flaviviridae virus, a Togaviridae virus, a Bornaviridae, a Filoviridae, a Paramyxoviridae, a Pneumoviridae, a Rhabdoviridae, an Arenaviridae, a Bunyaviridae, an Orthomyxoviridae, or a Deltavirus. In certain example embodiments, the virus is Coronavirus, SARS, Poliovirus, Rhinovirus, Hepatitis A, Norwalk virus, Yellow fever virus, West Nile virus, Hepatitis C virus, Dengue fever virus, Zika virus, Rubella virus, Ross River virus, Sindbis virus, Chikungunya virus, Borna disease virus, Ebola virus, Marburg virus, Measles virus, Mumps virus, Nipah virus, Hendra virus, Newcastle disease virus, Human respiratory syncytial virus, Rabies virus, Lassa virus, Hantavirus, Crimean-Congo hemorrhagic fever virus, Influenza, or Hepatitis D virus.

**[0104]** In certain example embodiments, the virus may be a retrovirus. Example retroviruses that may be detected using the embodiments disclosed herein include one or more of or any combination of viruses of the Genus Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus, Spumavirus, or the Family Metaviridae, Pseudoviridae, and Retroviridae (including HIV), Hepadnaviridae (including Hepatitis B virus), and Caulimoviridae (including Cauliflower mosaic virus).

**[0105]** In certain example embodiments, the virus is a DNA virus. Example DNA viruses that may be detected using the embodiments disclosed herein include one or more of (or any combination of) viruses from the Family Myoviridae, Podoviridae, Siphoviridae, Alloherpesviridae, Herpesviridae (including human herpes virus, and Varicella Zoster virus), Malocoherpesviridae, Lipothrixviridae, Rudiviridae, Adenoviridae, Ampullaviridae, Ascoviridae, Asfarviridae (including African swine fever virus), Baculoviridae,

Cicaudaviridae, Clavaviridae, Corticoviridae, Fuselloviridae, Globuloviridae, Guttaviridae, Hytrosaviridae, Iridoviridae, Maseilleviridae, Mimiviridae, Nudiviridae, Nimaviridae, Pandoraviridae, Papillomaviridae, Phycodnaviridae, Plasmaviridae, Polydnaviruses, Polyomaviridae (including Simian vims 40, JC virus, BK virus), Poxviridae (including Cowpox and smallpox), Sphaerolipoviridae, Tectiviridae, Turriviridae, Dinodnavirus, Salterprovirus, Rhizidovirus, among others.

**[0106]** In certain example embodiments, the target pathogen may be a protozoon. Examples of protozoa include without limitation any one or more of (or any combination of), Euglenozoa, Heterolobosea, Diplomonadida, Amoebozoa, Blastocystic, and Apicomplexa. Example Euglenoza include, but are not limited to, *Trypanosoma cruzi* (Chagas disease), *T. brucei gambiense, T. brucei rhodesiense, Leishmania braziliensis, L. infantum, L. mexicana, L. major, L. tropica,* and *L. donovani.* Example Heterolobosea include, but are not limited to, *Naegleria fowleri.* Example Diplomonadid include, but are not limited to, *Giardia intestinalis (G. lamblia, G. duodenalis).* Example Amoebozoa include, but are not limited to, *Acanthamoeba castellanii, Balamuthia madrillaris, Entamoeba histolytica.* Example Blastocystis include, but are not limited to, *Blastocystic hominis.* Example Apicomplexa include, but are not limited to, *Babesia microti, Cryptosporidium parvum, Cyclospora cayetanensis, Plasmodium falciparum, P. vivax, P. ovale, P. malariae,* and *Toxoplasma gondii.*

**[0107]** In certain example embodiments, the target epitopes are present on a commensal microorganism. The Human Microbiome Project sequenced the genome of the human microbiota, focusing particularly on the microbiota that normally inhabit the skin, mouth, nose, digestive tract, and vagina (see, e.g., hmpdacc.org/hmp/). MHCII epitopes may be predicted any of the microrganisms described.

**[0108]** The four dominant bacterial phyla in the human gut are Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria. Most bacteria belong to the genera Bacteroides, Clostridium, Faecalibacterium, Eubacterium, Ruminococcus, Peptococcus, Peptostreptococcus, and Bifidobacterium. Other genera, such as Escherichia and Lactobacillus, are present to a lesser extent. Species from the genus Bacteroides alone constitute about 30% of all bacteria in the gut, suggesting that this genus is especially important in the functioning of the host.

**[0109]** Fungal genera that have been detected in the gut include Candida, Saccharomyces, Aspergillus, Penicillium, Rhodotorula, Trametes, Pleospora, Sclerotinia, Bullera, and Galactomyces, among others. Rhodotorula is most frequently found in individuals with

inflammatory bowel disease while Candida is most frequently found in individuals with hepatitis B cirrhosis and chronic hepatitis B.

**[0110]** Archaea constitute another large class of gut flora which are important in the metabolism of the bacterial products of fermentation.

**[0111]** A number of types of bacteria, such as Actinomyces viscosus and A. naeslundii, live in the mouth.

**[0112]** The vaginal microflora consists mostly of various lactobacillus species (e.g., Lactobacillus acidophilus, L. iners, L. crispatus, L. jensenii, L. delbruekii and L. gasseri).

**[0113]** In certain example embodiments, the target epitopes are present on an allergen. As used herein the term "allergen" may refer to an antigen, microorganism, plant, or product thereof that produces an abnormal immune response in which the immune system fights off a perceived threat that would otherwise be harmless to the body. Allergens can be found in a variety of sources (e.g., animal products, foods, insects, mold spores, plants and chemicals). Allergens can include, but are not limited to dust mite, pollen, spores, poison ivy, poison oak, pet dander, royal jelly, peanuts (a legume), nuts, insect bites or stings, seafood and shellfish.

**[0114]** Method 200 begins at block 205, wherein a set of candidate antigens are generated from an input genome sequence. Turning to Figure 3, the process of generating a set of candidate peptides from one or more input genome sequences is described in more detail.

**[0115]** Method 205 begin at block 305, where a set of predicted gens is generated from one or more genome input sequences, such as from genomic sequence input files 107. Predicted genes may be identified using various gene prediction methodologies. The gene prediction may be *ab initio* including both signal-based and content-based approaches. In certain example embodiments, the gene prediction methodology is a hidden Markov model (HMM)-based method, or machine learning based, including neural network based approaches. Example gene prediction algorithms that may be used include, but are not limited to, GLIMMER, GeneMark, GENSCAN, geneid, SNAP, mSplicer, CONTRAST, or GENE. Methods that combine pattern recognition and machine learning such as Maker and Augustus may also be used. Other methods may include comparative genomic approaches, such as those employed by TWINSCAN, N-SCAN, and CONTRAST, may be used. Additional example gene prediction methods are disclosed in Mathe *et al.* Nucleic Acids Research (2002) 30(19):4103-41 17.

**[0116]** At block 310, the gene prediction module 111 then reduces the predicted gene set by selecting only those genes that are identified as producing protein products that are located on the cell surface or are secreted as extracellular proteins. An example method for predicting subcellular localization include PSORTb 3.0 as described in Yu *etal.* Bioinformatics 26, 1608-

1615 (2010). Other similar and known computer-based subcellular localization prediction methods may also be used.

**[0117]** At block 315, the peptide prediction module 111 then takes the sub-set of identified cell-surface and extracellular proteins and mask any intracellular regions and transmembrane domains in the identified subset. Methods for protein topology recognition that may be applied included those described in Tusnady *et al.* Journal of Molecular Biology 283:489-506 1998). Similar and other known topology prediction methods may be used. In certain example embodiments, the masked region may be expanded to include a flanking buffer region around the identify intracellular and transmembrane domains. The flanking region may comprise 1, 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 amino acid in either or both of the N- or C-terminal directions from the intracellular region or transmembrane domain. Not being bound by a theory, the flanking region may be differentially accessible based on specific antigens (e.g., antigens in the training set) and the flanking regions either increased or decreased.

**[0118]** At block 320, the peptide prediction module 111, then excludes from further analysis, regions of the candidate proteins that fall within small domains or between a series of domains. The metric used for domain mapping may include the distance to the upstream/downstream domains, the density of flanking domains and the size of the domain. In certain example embodiments, a small domain is a domain of 30 or fewer amino acids. In certain other example embodiments, a region of a candidate protein is excluded if it falls within 8 amino acids of a small domain. In other example embodiments, a region of a candidate protein is excluded if it falls within a series of domains. In certain embodiments, a region of a candidate protein is excluded if it falls within an interdomain region of 20 amino acids or less. Regions of the candidate proteins may be excluded based on one or more of the above criteria. The final result is a data file comprising a list of candidate antigens from each candidate protein. The candidate antigens are those portions of the candidate proteins that remain after all of the exclusion and masking steps above are completed. The candidate antigen data file may be stored in the candidate antigen library 112. The method then returns to block 210 of FIG. 2.

**[0119]** At block 210 of FIG. 2, the epitope ranking module 113 then takes the candidate antigens as input and proceeds to generate a list of ranked candidate epitopes. Many factors contribute to the score for ranking (e.g., MHC binding, cellular localization, protein topology, domain architecture). The process is described in more detail with reference to FIG. 4

**[0120]** Method 210 begins at block 405 of FIG. 4, wherein the epitope ranking module 113 encodes each amino acid of each candidate antigen into a p-dimensional binary vector b with half of the vector elements being 1 and the rest being 0, chosen at random.

**[0121]** At block 410, the epitope ranking module 113 coverts each a p-dimensional binary vector b into a descriptor matrix S. Therefore, given a peptide with length 1longer than *k*, it is first converted to an *I x b* descriptor matrix *S*, in which *Sy* = 1 if the *i-th* amino acid's 1-valued indices overlap with*j* , otherwise *Sy* = 0. The matrix *S* is then normalized by row sum to become *S'* such that

$$S'_{ij} = S_{ij} / \sum_{j=1}^{b} S_{ij}$$

**[0122]** At block 415, the epitope ranking module 113 then convolutes *S'* into a *(l-k+l) x d* matrix *X*, where *d* is the number of pre-trained motif network models within the overall model and *k* is the length of such binding cores. *Xy* represents the score of motif network model *j* aligned to position *i*. The cohort of motif network models is arranged in a *d x k x b* array *H* with $Hy_k$ being the *d-th* motif network model aligned to the *k-th* position of the *b-th* set of amino acids. This sequence conversion and convolution setup is similar to model developed by Alipanahi et al. (Alipanahi et al., 2015). Methods for generating pre-trained motif network models are described in further detail below

**[0123]** At block 420, the epitope ranking module 113 then transforms the scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for neural network analysis and ranking of epitopes. The output of the neural network is a list of scored epitopes. In certain example embodiments, the epitopes may be ranked by immunodominance. In certain example embodiments, the convoluted matrix *X*, it is filtered with a max-rectified linear unit (ReLU) layer, and the rectified matrix *Y* is then fed into a max pooling stage to be transformed into d-dimensional vector Z, in which,

$$z_j = max(Y_{1j}, Y_{2j}, \ldots, Y_{nj})$$

This d-dimensional vector Z is then used as input for a neural network with a Maxout dropout model. Z is then used as the input for a standard output layer for final prediction calculation. While not being bound by the following theory, the method minimizes the prediction error as measured 1 by 1-norm distance of all the peptides. In one embodiment, the back propagation with stochastic gradient descent (SGD) method using mini batch size at 64 may be used to

reach optimal weights. To train the weights, mouse epitope peptides may be used as a training set of true positive peptides. *In silico* true negative peptides may also be used as a training set by surveying an equal number of peptides that are not part of the peptidomics data readout at random.

**[0124]** In one embodiment, a neural network is trained by a method comprising exposing dendritic or antigen presenting cells to a pool of potential binding antigens. In certain embodiments, the pool includes non-self-antigens. In preferred embodiments, the pool of potential antigens is derived from a target pathogen or cell type. Not being bound by a theory, exposure of the dendritic or antigen presenting cells to the target pathogen or cell type allows for the determination of improved binding parameters for MHCII molecules because non-self-binding peptides may bind with higher affinity. The dendritic cells may be lysed and MHCII peptide complexes may be isolated using an affinity reagent specific for the MHCII molecule. In preferred embodiments, antibodies are used. The antibodies may be bound to a solid support, such as, but not limited to magnetic beads or protein A or G beads. Upon isolation of the MHCII peptide complexes, the bound peptides may be determined by protein sequencing, such as, but not limited to mass spectrometry or Edman degradation. The peptides may include peptides originating from the target pathogen or cell type and may also include peptides originating from the dendritic cell or antigen presenting host organism. The dendritic cells or antigen presenting cells may originate from a mammalian host organism, such as, but not limited to a human or mouse host organism. The bound peptides may be used for generating detectors in block 230 by testing randomly initialized parameters for the peptides. Such parameters may include, but are not limited to, amino acid sequence or net charge of peptide. The parameters may be tested in a three-fold cross validation scheme. Sets of randomly selected parameters may be tested and the best performing set provided to the deep neural network. In certain embodiments, more than 10, 20, 30, or 40, preferably about 30 parameter sets are constructed. The sets may include more than 10, 20, 30, or 40, preferably about 30 parameters. Based on the parameters most consistent with the identified peptides a set is selected. The optimal parameters may be identified by having the optimal ROC AUC score. The selected training parameters are provided to the deep neural network 116 for predicting unknown binding peptides from a genome sequence. FIG. 6 provides an example experimental work flow for identifying peptides which may then be used to train neural networks in accordance with the embodiments disclosed herein. The experimental work flow is described in greater detail in the Examples below.

**Other Example Computer Embodiments**

**[0125]** FIG. 5 depicts a computing machine 2000 and a module 2050 in accordance with certain example embodiments. The computing machine 2000 may correspond to any of the various computers, servers, mobile devices, embedded systems, or computing systems presented herein. The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 in performing the various methods and processing functions presented herein. The computing machine 2000 may include various internal or attached components such as a processor 2010, system bus 2020, system memory 2030, storage media 2040, input/output interface 2060, and a network interface 2070 for communicating with a network 2080.

**[0126]** The computing machine 2000 may be implemented as a conventional computer system, an embedded controller, a laptop, a server, a mobile device, a smartphone, a set-top box, a kiosk, a vehicular information system, one more processors associated with a television, a customized machine, any other hardware platform, or any combination or multiplicity thereof. The computing machine 2000 may be a distributed system configured to function using multiple computing machines interconnected via a data network or bus system.

**[0127]** The processor 2010 may be configured to execute code or instructions to perform the operations and functionality described herein, manage request flow and address mappings, and to perform calculations and generate commands. The processor 2010 may be configured to monitor and control the operation of the components in the computing machine 2000. The processor 2010 may be a general purpose processor, a processor core, a multiprocessor, a reconfigurable processor, a microcontroller, a digital signal processor ("DSP"), an application specific integrated circuit ("ASIC"), a graphics processing unit ("GPU"), a field programmable gate array ("FPGA"), a programmable logic device ("PLD"), a controller, a state machine, gated logic, discrete hardware components, any other processing unit, or any combination or multiplicity thereof. The processor 2010 may be a single processing unit, multiple processing units, a single processing core, multiple processing cores, special purpose processing cores, co-processors, or any combination thereof. According to certain embodiments, the processor 2010 along with other components of the computing machine 2000 may be a virtualized computing machine executing within one or more other computing machines.

**[0128]** The system memory 2030 may include non-volatile memories such as read-only memory ("ROM"), programmable read-only memory ("PROM"), erasable programmable read-only memory ("EPROM"), flash memory, or any other device capable of storing program instructions or data with or without applied power. The system memory 2030 may also include volatile memories such as random access memory ("RAM"), static random access memory

("SRAM"), dynamic random access memory ("DRAM"), and synchronous dynamic random access memory ("SDRAM"). Other types of RAM also may be used to implement the system memory 2030. The system memory 2030 may be implemented using a single memory module or multiple memory modules. While the system memory 2030 is depicted as being part of the computing machine 2000, one skilled in the art will recognize that the system memory 2030 may be separate from the computing machine 2000 without departing from the scope of the subject technology. It should also be appreciated that the system memory 2030 may include, or operate in conjunction with, a non-volatile storage device such as the storage media 2040.

**[0129]** The storage media 2040 may include a hard disk, a floppy disk, a compact disc read only memory ("CD-ROM"), a digital versatile disc ("DVD"), a Blu-ray disc, a magnetic tape, a flash memory, other non-volatile memory device, a solid state drive ("SSD"), any magnetic storage device, any optical storage device, any electrical storage device, any semiconductor storage device, any physical-based storage device, any other data storage device, or any combination or multiplicity thereof. The storage media 2040 may store one or more operating systems, application programs and program modules such as module 2050, data, or any other information. The storage media 2040 may be part of, or connected to, the computing machine 2000. The storage media 2040 may also be part of one or more other computing machines that are in communication with the computing machine 2000 such as servers, database servers, cloud storage, network attached storage, and so forth.

**[0130]** The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 with performing the various methods and processing functions presented herein. The module 2050 may include one or more sequences of instructions stored as software or firmware in association with the system memory 2030, the storage media 2040, or both. The storage media 2040 may therefore represent examples of machine or computer readable media on which instructions or code may be stored for execution by the processor 2010. Machine or computer readable media may generally refer to any medium or media used to provide instructions to the processor 2010. Such machine or computer readable media associated with the module 2050 may comprise a computer software product. It should be appreciated that a computer software product comprising the module 2050 may also be associated with one or more processes or methods for delivering the module 2050 to the computing machine 2000 via the network 2080, any signal-bearing medium, or any other communication or delivery technology. The module 2050 may also comprise hardware circuits or information for configuring hardware circuits such as microcode or configuration information for an FPGA or other PLD.

**[0131]** The input/output ("I/O") interface 2060 may be configured to couple to one or more external devices, to receive data from the one or more external devices, and to send data to the one or more external devices. Such external devices along with the various internal devices may also be known as peripheral devices. The I/O interface 2060 may include both electrical and physical connections for operably coupling the various peripheral devices to the computing machine 2000 or the processor 2010. The I/O interface 2060 may be configured to communicate data, addresses, and control signals between the peripheral devices, the computing machine 2000, or the processor 2010. The I/O interface 2060 may be configured to implement any standard interface, such as small computer system interface ("SCSI"), serial-attached SCSI ("SAS"), fiber channel, peripheral component interconnect ("PCI"), PCI express (PCIe), serial bus, parallel bus, advanced technology attached ("ATA"), serial ATA ("SATA"), universal serial bus ("USB"), Thunderbolt, FireWire, various video buses, and the like. The I/O interface 2060 may be configured to implement only one interface or bus technology. Alternatively, the I/O interface 2060 may be configured to implement multiple interfaces or bus technologies. The I/O interface 2060 may be configured as part of, all of, or to operate in conjunction with, the system bus 2020. The I/O interface 2060 may include one or more buffers for buffering transmissions between one or more external devices, internal devices, the computing machine 2000, or the processor 2010.

**[0132]** The I/O interface 2060 may couple the computing machine 2000 to various input devices including mice, touch-screens, scanners, biometric readers, electronic digitizers, sensors, receivers, touchpads, trackballs, cameras, microphones, keyboards, any other pointing devices, or any combinations thereof. The I/O interface 2060 may couple the computing machine 2000 to various output devices including video displays, speakers, printers, projectors, tactile feedback devices, automation control, robotic components, actuators, motors, fans, solenoids, valves, pumps, transmitters, signal emitters, lights, and so forth.

**[0133]** The computing machine 2000 may operate in a networked environment using logical connections through the network interface 2070 to one or more other systems or computing machines across the network 2080. The network 2080 may include wide area networks (WAN), local area networks (LAN), intranets, the Internet, wireless access networks, wired networks, mobile networks, telephone networks, optical networks, or combinations thereof. The network 2080 may be packet switched, circuit switched, of any topology, and may use any communication protocol. Communication links within the network 2080 may involve various digital or an analog communication media such as fiber optic cables, free-space

optics, waveguides, electrical conductors, wireless links, antennas, radio-frequency communications, and so forth.

**[0134]** The processor 2010 may be connected to the other elements of the computing machine 2000 or the various peripherals discussed herein through the system bus 2020. It should be appreciated that the system bus 2020 may be within the processor 2010, outside the processor 2010, or both. According to some embodiments, any of the processor 2010, the other elements of the computing machine 2000, or the various peripherals discussed herein may be integrated into a single device such as a system on chip ("SOC"), system on package ("SOP"), or ASIC device.

**[0135]** Embodiments may comprise a computer program that embodies the functions described and illustrated herein, wherein the computer program is implemented in a computer system that comprises instructions stored in a machine-readable medium and a processor that executes the instructions. However, it should be apparent that there could be many different ways of implementing embodiments in computer programming, and the embodiments should not be construed as limited to any one set of computer program instructions. Further, a skilled programmer would be able to write such a computer program to implement an embodiment of the disclosed embodiments based on the appended flow charts and associated description in the application text. Therefore, disclosure of a particular set of program code instructions is not considered necessary for an adequate understanding of how to make and use embodiments. Further, those skilled in the art will appreciate that one or more aspects of embodiments described herein may be performed by hardware, software, or a combination thereof, as may be embodied in one or more computing systems. Moreover, any reference to an act being performed by a computer should not be construed as being performed by a single computer as more than one computer may perform the act.

**[0136]** The example embodiments described herein can be used with computer hardware and software that perform the methods and processing functions described herein. The systems, methods, and procedures described herein can be embodied in a programmable computer, computer-executable software, or digital circuitry. The software can be stored on computer-readable media. For example, computer-readable media can include a floppy disk, RAM, ROM, hard disk, removable media, flash memory, memory stick, optical media, magneto-optical media, CD-ROM, etc. Digital circuitry can include integrated circuits, gate arrays, building block logic, field programmable gate arrays (FPGA), etc.

**Immunogenic compositions**

**[0137]** The present invention provides for predicting MHCII binding epitopes useful in formulating an immunogenic composition. The immunogenic composition may be used in the treatment of a patient in need thereof. The immunogenic composition may be a protective or tolerizing vaccine. In preferred embodiments, the immunogenic composition activates CD4+ T cells to generate a humoral immune response. In certain embodiments, the peptides enhance a CD8+ CTL response. A tolerizing vaccine may be formulated with antigenic epitopes specific for an allergen or for an autoimmunity antigen. A protective vaccine may be formulated with antigenic epitopes specific for a pathogen or for a cancer cell.

**[0138]** The terms "subject," "individual," and "patient" are used interchangeably herein to refer to a vertebrate, preferably a mammal, more preferably a human. Mammals include, but are not limited to, murines, simians, humans, farm animals, sport animals, and pets. Tissues, cells and their progeny of a biological entity obtained *in vivo* or cultured *in vitro* are also encompassed.

**[0139]** The terms "therapeutic agent", "therapeutic capable agent" or "treatment agent" are used interchangeably and refer to a molecule or compound that confers some beneficial effect upon administration to a subject. The beneficial effect includes enablement of diagnostic determinations; amelioration of a disease, symptom, disorder, or pathological condition; reducing or preventing the onset of a disease, symptom, disorder or condition; and generally counteracting a disease, symptom, disorder or pathological condition.

**[0140]** As used herein, "treatment" or "treating," or "palliating" or "ameliorating" are used interchangeably. These terms refer to an approach for obtaining beneficial or desired results including but not limited to a therapeutic benefit and/or a prophylactic benefit. By therapeutic benefit is meant any therapeutically relevant improvement in or effect on one or more diseases, conditions, or symptoms under treatment. For prophylactic benefit, the compositions may be administered to a subject at risk of developing a particular disease, condition, or symptom, or to a subject reporting one or more of the physiological symptoms of a disease, even though the disease, condition, or symptom may not have yet been manifested.

**[0141]** The term "effective amount" or "therapeutically effective amount" refers to the amount of an agent that is sufficient to effect beneficial or desired results. The therapeutically effective amount may vary depending upon one or more of: the subject and disease condition being treated, the weight and age of the subject, the severity of the disease condition, the manner of administration and the like, which can readily be determined by one of ordinary skill in the art. The term also applies to a dose that will provide an image for detection by any one of the imaging methods described herein. The specific dose may vary depending on one or

more of: the particular agent chosen, the dosing regimen to be followed, whether it is administered in combination with other compounds, timing of administration, the tissue to be imaged, and the physical delivery system in which it is carried.

**[0142]** The immune system can be classified into two functional subsystems: the innate and the acquired immune system. The innate immune system is the first line of defense against infections, and most potential pathogens are rapidly neutralized by this system before they can cause, for example, a noticeable infection. The acquired immune system reacts to molecular structures, referred to as antigens, of the intruding organism. There are two types of acquired immune reactions, which include the humoral immune reaction and the cell-mediated immune reaction. In the humoral immune reaction, antibodies secreted by B cells into bodily fluids bind to pathogen-derived antigens, leading to the elimination of the pathogen through a variety of mechanisms, e.g. complement-mediated lysis. In the cell-mediated immune reaction, T-cells capable of destroying other cells are activated. For example, if proteins associated with a disease are present in a cell, they are fragmented proteolytically to peptides within the cell. Specific cell proteins then attach themselves to the antigen or peptide formed in this manner and transport them to the surface of the cell, where they are presented to the molecular defense mechanisms, in particular T-cells, of the body. Cytotoxic T cells recognize these antigens and kill the cells that harbor the antigens.

**[0143]** The molecules that transport and present peptides on the cell surface are referred to as proteins of the major histocompatibility complex (MHC). MHC proteins are classified into two types, referred to as MHC class I and MHC class II. The structures of the proteins of the two MHC classes are very similar; however, they have very different functions. Proteins of MHC class I are present on the surface of almost all cells of the body, including most tumor cells. MHC class I proteins are loaded with antigens that usually originate from endogenous proteins or from pathogens present inside cells, and are then presented to naive or cytotoxic T-lymphocytes (CTLs). MHC class II proteins are present on dendritic cells (DC), B-lymphocytes, macrophages and other antigen-presenting cells (APC). They mainly present peptides, which are processed from external antigen sources, i.e. outside of the cells, to T-helper (Th) cells.

**[0144]** Most of the peptides bound by the MHC class I proteins originate from cytoplasmic proteins produced in the healthy host cells of an organism itself, and do not normally stimulate an immune reaction. Accordingly, cytotoxic T-lymphocytes that recognize such self-peptide-presenting MHC molecules of class I are deleted in the thymus (central tolerance) or, after their release from the thymus, are deleted or inactivated, i.e. tolerized (peripheral tolerance). MHC

molecules are capable of stimulating an immune reaction when they present peptides to non-tolerized T-lymphocytes. Cytotoxic T-lymphocytes have both T-cell receptors (TCR) and CD8 molecules on their surface. T-Cell receptors are capable of recognizing and binding peptides complexed with the molecules of MHC class I. Each cytotoxic T-lymphocyte expresses a unique T-cell receptor which is capable of binding specific MHC/peptide complexes.

**[0145]** MHC Class II molecules are required for the activation of CD4+ T cells. CD4+ T cells depend mainly on infiltrating APCs that either pick up available antigens or engulf tumor cells or pathogens. The MHC Class II antigen presentation pathway can be targeted in an immunotherapy (see e.g., Thibodeau, et al., Targeting the MHC Class II antigen presentation pathway in cancer immunotherapy. Oncoimmunology. 2012 Sep 1;1(6):908-916). Antigens can gain access to MHC Class II-loading compartments by multiple distinct means. For example, transmembrane proteins from the plasma membrane can be endocytosed and sent to lysosomes for degradation. Phagocytosis is also a mechanism antigens can gain access to MHC Class II loading compartments. For example, when a macrophage ingests a pathogenic microorganism, or cancer cell the pathogen or cancer cell becomes trapped in a phagosome which then fuses with a lysosome to form a phagolysosome. Within the phagolysosome, enzymes and toxic peroxides digest the pathogen or cancer cell. Cytoplasmic and nuclear antigens can also be engulfed by autophagy.

**[0146]** In certain embodiments, a subject is administered an immunogenic composition comprising the antigenic epitopes of the present invention. In one embodiment, the peptides of the present invention are administered to a subject in need thereof.

**[0147]** In one embodiment, the immunogenic composition comprises natural or artificial APCs that have been manipulated *in vitro* to display the antigens of the present invention. DCs are potent antigen-presenting cells that initiate T cell immunity and can be used as vaccines when loaded with one or more peptides of interest, for example, by direct peptide injection. The dendritic cells may be isolated from peripheral blood by means of leukapheresis. They may be cultured *in vitro* with the cytokines GM-CSF and interleukin-4, loaded with antigen, and matured *ex vivo* to enhance antigen presentation and costimulation of T-cells before being injected into patients. Antigen may be delivered to DCs as peptide, protein, or RNA. B cells may also serve as efficient APCs for the expansion of antigen specific CD4+ T cells. B cells can present MHC Class II epitopes when pulsed exogenously and can also promote MHC Class I cross-presentation. Antigenic epitopes may also be presented by acellular artificial APCs that consist of microbeads, liposomes or exosomes.

**[0148]** It is contemplated within the scope of the invention that antigen loaded DCs may be prepared using the synthetic TLR 3 agonist Polyinosinic-Polycytidylic Acid-poly-L-lysine Carboxymethylcellulose (Poly-ICLC) to stimulate the DCs. Poly-ICLC is a potent individual maturation stimulus for human DCs as assessed by an upregulation of CD83 and CD86, induction of interleukin-12 (IL-12), tumor necrosis factor (TNF), interferon gamma-induced protein 10 (IP-10), interleukin 1 (IL-1), and type I interferons (IFN), and minimal interleukin 10 (IL-10) production. DCs may be differentiated from frozen peripheral blood mononuclear cells (PBMCs) obtained by leukapheresis, while PBMCs may be isolated by Ficoll gradient centrifugation and frozen in aliquots.

**[0149]** Illustratively, the following 7 day activation protocol may be used. Day 1—PBMCs are thawed and plated onto tissue culture flasks to select for monocytes which adhere to the plastic surface after 1-2 hr incubation at 37°C in the tissue culture incubator. After incubation, the lymphocytes are washed off and the adherent monocytes are cultured for 5 days in the presence of interleukin-4 (IL-4) and granulocyte macrophage-colony stimulating factor (GM-CSF) to differentiate to immature DCs. On Day 6, immature DCs are pulsed with the keyhole limpet hemocyanin (KLH) protein which serves as a control for the quality of the vaccine and may boost the immunogenicity of the vaccine. The DCs are stimulated to mature, loaded with peptide antigens, and incubated overnight. On Day 7, the cells are washed, and frozen in 1 ml aliquots containing 4-20 x $10^6$ cells using a controlled-rate freezer. Lot release testing for the batches of DCs may be performed to meet minimum specifications before the DCs are injected into patients (see e.g., Sabado et al. (2013) Preparation of tumor antigen-loaded mature dendritic cells for immunotherapy, J. Vis Exp. Aug 1;(78). doi: 10.3791/50085).

**[0150]** Thus, in one embodiment of the present invention the vaccine or immunogenic composition containing at least one antigen presenting cell is pulsed or loaded with one or more peptides of the present invention. Alternatively, peripheral blood mononuclear cells (PBMCs) isolated from a patient may be loaded with peptides *ex vivo* and injected back into the patient. As an alternative the antigen presenting cell comprises an expression construct encoding a peptide of the present invention. The polynucleotide may be any suitable polynucleotide and it is preferred that it is capable of transducing the dendritic cell, thus resulting in the presentation of a peptide and induction of immunity.

**[0151]** In another embodiment, the immunogenic composition comprises recombinant antigens coupled to monoclonal antibodies directed against DC surface receptors.

**[0152]** Amounts effective for an immunogenic composition can depend on, e.g., the peptide 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, but generally range for the initial immunization (that is for therapeutic or prophylactic administration) from about 1.0 μg to about 50,000 μg of peptide for a 70 kg patient, followed by boosting dosages or from about 1.0 μg to about 10,000 μg of peptide pursuant to a boosting regimen over weeks to months depending upon the patient's response and condition.

**[0153]** The pharmaceutical compositions (e.g., vaccine compositions) for therapeutic treatment are intended for parenteral, topical, nasal, oral or local administration. Preferably, the pharmaceutical compositions are administered parenterally, e.g., intravenously, subcutaneously, intradermally, or intramuscularly. The compositions may be administered at the site of surgical excision to induce a local immune response to the tumor. The invention provides compositions for parenteral administration which comprise a solution of the peptides and vaccine or immunogenic compositions are dissolved or suspended in an acceptable carrier, preferably an aqueous carrier. A variety of aqueous carriers may be used, e.g., water, buffered water, 0.9% saline, 0.3% glycine, hyaluronic acid and the like. These compositions may be sterilized by conventional, well known sterilization techniques, or may be sterile filtered. The resulting aqueous solutions may 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.

**[0154]** A liposome suspension containing a peptide may 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. For targeting to the immune cells, a ligand, such as, e.g., antibodies or fragments thereof specific for cell surface determinants of the desired immune system cells, can be incorporated into the liposome.

**[0155]** For solid compositions, conventional or nanoparticle nontoxic solid carriers may be used which include, for example, pharmaceutical grades of mannitol, lactose, starch, magnesium stearate, sodium saccharin, talcum, cellulose, glucose, sucrose, magnesium carbonate, and the like. For oral administration, a pharmaceutically acceptable nontoxic composition is formed by incorporating any of the normally employed excipients, such as those

carriers previously listed, and generally 10-95% of active ingredient, that is, one or more peptides of the invention, and more preferably at a concentration of 25%-75%.

**[0156]** For aerosol administration, the immunogenic peptides are preferably supplied in finely divided form along with a surfactant and propellant. Typical percentages of peptides are 0.01 %-20% by weight, preferably 1%-10%. The surfactant can, of course, be nontoxic, and preferably soluble in the propellant. Representative of such agents are the esters or partial esters of fatty acids containing from 6 to 22 carbon atoms, such as caproic, octanoic, lauric, palmitic, stearic, linoleic, linolenic, olesteric and oleic acids with an aliphatic polyhydric alcohol or its cyclic anhydride. Mixed esters, such as mixed or natural glycerides may be employed. The surfactant may constitute 0.1%-20% by weight of the composition, preferably 0.25-5%. The balance of the composition is ordinarily propellant. A carrier can also be included as desired, as with, e.g., lecithin for intranasal delivery.

**[0157]** As described in further detail herein, *in vitro* production of antigenic peptides may occur by a variety of methods known to one of skill in the art such as, for example, peptide synthesis or expression of a peptide/polypeptide from a DNA or RNA molecule in any of a variety of bacterial, eukaryotic, or viral recombinant expression systems, followed by purification of the expressed peptide/polypeptide. Alternatively, antigenic peptides may be produced *in vivo* by introducing molecules (e.g., DNA, RNA, viral expression systems, and the like) that encode antigenic peptides into a subject, whereupon the encoded antigenic peptides are expressed.

**[0158]** Peptides can be readily synthesized chemically utilizing reagents that are free of contaminating bacterial or animal substances (Merrifield RB: Solid phase peptide synthesis. I. The synthesis of atetrapeptide. J. Am. Chem. Soc. 85:2149-54, 1963). In certain embodiments, antigenic peptides are prepared by (1) parallel solid-phase synthesis on multi-channel instruments using uniform synthesis and cleavage conditions; (2) purification over aRP-HPLC column with column stripping; and re-washing, but not replacement, between peptides; followed by (3) analysis with a limited set of the most informative assays. Synthetic peptides provide a particularly useful means to prepare multiple immunogens efficiently and to rapidly translate identification of mutant epitopes to an effective vaccine or immunogenic composition. Peptides can be readily synthesized chemically and easily purified utilizing reagents free of contaminating bacteria or animal substances. The small size allows a clear focus on the mutated region of the protein and also reduces irrelevant antigenic competition from other components (unmutated protein or viral vector antigens).

**[0159]** In one embodiment, the drug formulation is a multi-epitope vaccine or immunogenic composition of long peptides. Such "long" peptides undergo efficient internalization, processing and cross-presentation in professional antigen-presenting cells such as dendritic cells, and have been shown to induce CTLs in humans (Melief and van der Burg, Immunotherapy of established (pre) malignant disease by synthetic long peptide vaccines Nature Rev Cancer 8:351 (2008)). In one embodiment, at least 1 peptide is prepared for immunization. In a preferred embodiment 20 or more peptides are prepared for immunization. In one embodiment, the neoantigenic peptide ranges from about 5 to about 50 amino acids in length. In another embodiment peptides from about 15 to about 35 amino acids in length is synthesized. In preferred embodiments, the peptide ranges from about 20 to about 35 amino acids in length.

**[0160]** The vaccine or immunogenic composition can further comprise an adjuvant and/or a carrier. Examples of useful adjuvants and carriers are given herein. The peptides and/or polypeptides in the 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.

**Vaccine or Immunogenic Composition Adjuvant**

**[0161]** Adjuvants are any substance whose admixture into the vaccine or immunogenic composition increases or otherwise modifies the immune response to the mutant peptide. Carriers are scaffold structures, for example a polypeptide or a polysaccharide, to which the neoantigenic peptides, is capable of being associated. Optionally, adjuvants are conjugated covalently or non-covalently to the peptides or polypeptides of the invention.

**[0162]** The ability of an adjuvant to increase the immune response to an antigen is typically manifested by a significant increase in 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 Th2 response into a primarily cellular, or Thl response.

**[0163]** Effective vaccine or immunogenic compositions advantageously include a strong adjuvant to initiate an immune response. As described herein, poly-ICLC, an agonist of TLR3 and the RNA helicase -domains of MDA5 and RIG3, has shown several desirable properties for a vaccine or immunogenic composition adjuvant. These properties include the induction of local and systemic activation of immune cells *in vivo*, production of stimulatory chemokines

and cytokines, and stimulation of antigen-presentation by DCs. Furthermore, poly-ICLC can induce durable CD4+ and CD8+ responses in humans. Importantly, striking similarities in the up regulation of transcriptional and signal transduction pathways were seen in subjects vaccinated with poly-ICLC and in volunteers who had received the highly effective, replication-competent yellow fever vaccine. Furthermore, >90% of ovarian carcinoma patients immunized with poly-ICLC in combination with a NY-ESO-1 peptide vaccine (in addition to Montanide) showed induction of CD4+ and CD8+ T cell, as well as antibody responses to the peptide in a recent phase 1 study. At the same time, poly-ICLC has been extensively tested in more than 25 clinical trials to date and exhibited a relatively benign toxicity profile. In addition to a powerful and specific immunogen the neoantigen peptides may be combined with an adjuvant (e.g., poly-ICLC) or another anti-neoplastic agent. Without being bound by theory, these neoantigens are expected to bypass central thymic tolerance (thus allowing stronger anti-tumor T cell response), while reducing the potential for autoimmunity (e.g., by avoiding targeting of normal self-antigens). An effective immune response advantageously includes a strong adjuvant to activate the immune system (Speiser and Romero, Molecularly defined vaccines for cancer immunotherapy, and protective T cell immunity Seminars in Immunol 22: 144 (2010)). For example, Toll-like receptors (TLRs) have emerged as powerful sensors of microbial and viral pathogen "danger signals", effectively inducing the innate immune system, and in turn, the adaptive immune system (Bhardwaj and Gnjatic, TLR AGONISTS: Are They Good Adjuvants? Cancer J. 16:382-391 (2010)). Among the TLR agonists, poly-ICLC (a synthetic double-stranded RNA mimic) is one of the most potent activators of myeloid-derived dendritic cells. In a human volunteer study, poly-ICLC has been shown to be safe and to induce a gene expression profile in peripheral blood cells comparable to that induced by one of the most potent live attenuated viral vaccines, the yellow fever vaccine YF-17D (Caskey et al, Synthetic double-stranded RNA induces innate immune responses similar to a live viral vaccine in humans J Exp Med 208:2357 (201 1)). In a preferred embodiment Hiltonol®, a GMP preparation of poly-ICLC prepared by Oncovir, Inc, is utilized as the adjuvant. In other embodiments, other adjuvants described herein are envisioned. For instance oil-in-water, water-in-oil or multiphasic W/O/W; see, e.g., US 7,608,279 and Aucouturier et al, Vaccine 19 (2001), 2666-2672, and documents cited therein.

**[0164]** Adjuvants are any substance whose admixture into the vaccine or immunogenic composition increases or otherwise modifies the immune response to the mutant peptide. Carriers are scaffold structures, for example a polypeptide or a polysaccharide, to which the

neoantigenic peptides, is capable of being associated. Optionally, adjuvants are conjugated covalently or non-covalently to the peptides or polypeptides of the invention.

**[0165]** The ability of an adjuvant to increase the immune response to an antigen is typically manifested by a significant increase in 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 Th2 response into a primarily cellular, or Thl response.

**[0166]** Suitable adjuvants include, but are not limited to 1018 ISS, aluminum salts, Amplivax, AS15, BCG, CP-870,893, CpG7909, CyaA, dSLIM, GM-CSF, IC30, IC31, Imiquimod, ImuFact IMP321, IS Patch, ISS, ISCOMATRIX, Juvlmmune, 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. 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; 92:3-1 1). Also cytokines may 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).

**[0167]** Toll like receptors (TLRs) may also be used as adjuvants, and are important members of the family of pattern recognition receptors (PRRs) which recognize conserved motifs shared by many micro-organisms, termed "pathogen-associated molecular patterns" (PAMPS). Recognition of these "danger signals" activates multiple elements of the innate and adaptive immune system. TLRs are expressed by cells of the innate and adaptive immune systems such as dendritic cells (DCs), macrophages, T and B cells, mast cells, and granulocytes and are localized in different cellular compartments, such as the plasma membrane, lysosomes,

endosomes, and endolysosomes. Different TLRs recognize distinct PAMPS. For example, TLR4 is activated by LPS contained in bacterial cell walls, TLR9 is activated by unmethylated bacterial or viral CpG DNA, and TLR3 is activated by double stranded RNA. TLR ligand binding leads to the activation of one or more intracellular signaling pathways, ultimately resulting in the production of many key molecules associated with inflammation and immunity (particularly the transcription factor NF-κB and the Type-I interferons). TLR mediated DC activation leads to enhanced DC activation, phagocytosis, upregulation of activation and co-stimulation markers such as CD80, CD83, and CD86, expression of CCR7 allowing migration of DC to draining lymph nodes and facilitating antigen presentation to T cells, as well as increased secretion of cytokines such as type I interferons, IL-12, and IL-6. All of these downstream events are critical for the induction of an adaptive immune response.

**[0168]** Among the most promising cancer vaccine or immunogenic composition adjuvants currently in clinical development are the TLR9 agonist CpG and the synthetic double-stranded RNA (dsRNA) TLR3 ligand poly-ICLC. In preclinical studies poly-ICLC appears to be the most potent TLR adjuvant when compared to LPS and CpG due to its induction of pro-inflammatory cytokines and lack of stimulation of IL-10, as well as maintenance of high levels of co-stimulatory molecules in DCsl. Furthermore, poly-ICLC was recently directly compared to CpG in non-human primates (rhesus macaques) as adjuvant for a protein vaccine or immunogenic composition consisting of human papillomavirus (HPV)16 capsomers (Stahl-Hennig C, Eisenblatter M, Jasny E, et al. Synthetic double-stranded RNAs are adjuvants for the induction of T helper 1 and humoral immune responses to human papillomavirus in rhesus macaques. PLoS pathogens. Apr 2009;5(4)).

**[0169]** CpG immuno stimulatory oligonucleotides have also been reported to enhance the effects of adjuvants in a vaccine or immunogenic composition setting. Without being bound by theory, CpG oligonucleotides act by activating the innate (non- adaptive) immune system via Toll-like receptors (TLR), mainly TLR9. CpG triggered TLR9 activation enhances antigen- specific humoral and cellular responses to a wide variety of antigens, including peptide or protein antigens, live or killed viruses, dendritic cell vaccines, autologous cellular vaccines and polysaccharide conjugates in both prophylactic and therapeutic vaccines. More importantly, it enhances dendritic cell maturation and differentiation, resulting in enhanced activation of Thl cells and strong cytotoxic T- lymphocyte (CTL) generation, even in the absence of CD4 T-cell help. The Thl bias induced by TLR9 stimulation is maintained even in the presence of vaccine adjuvants such as alum or incomplete Freund's adjuvant (IFA) that normally promote a Th2 bias. CpG oligonucleotides show even greater adjuvant activity when

formulated or co-administered with other adjuvants or in formulations such as microparticles, nano particles, lipid emulsions or similar formulations, which are especially necessary for inducing a strong response when the antigen is relatively weak. They also accelerate the immune response and enabled the antigen doses to be reduced by approximately two orders of magnitude, with comparable antibody responses to the full-dose vaccine without CpG in some experiments (Arthur M. Krieg, Nature Reviews, Drug Discovery, 5, Jun. 2006, 471-484). U.S. Pat. No. 6,406,705 Bl describes the combined use of CpG oligonucleotides, non-nucleic acid adjuvants and an antigen to induce an antigen- specific immune response. A commercially available CpG TLR9 antagonist is dSLIM (double Stem Loop Immunomodulator) by Mologen (Berlin, GERMANY), which is a preferred component of the pharmaceutical composition of the present invention. Other TLR binding molecules such as RNA binding TLR 7, TLR 8 and/or TLR 9 may also be used.

**[0170]** Other examples of useful adjuvants include, but are not limited to, chemically modified CpGs (e.g. CpR, Idera), Poly(I:C)(e.g. polyi:CI2U), non-CpG bacterial DNA or RNA as well as immunoactive small molecules and antibodies such as cyclophosphamide, sunitinib, bevacizumab, celebrex, NCX-4016, sildenafil, tadalafil, vardenafil, sorafinib, XL-999, CP- 547632, pazopanib, ZD2171, AZD2171, ipilimumab, tremelimumab, and SC58175, which may act therapeutically and/or as an adjuvant. The amounts and concentrations of adjuvants and additives useful in the context of the present invention 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).

**[0171]** Poly-ICLC is a synthetically prepared double-stranded RNA consisting of polyl and polyC strands of average length of about 5000 nucleotides, which has been stabilized to thermal denaturation and hydrolysis by serum nucleases by the addition of polylysine and carboxymethylcellulose. The compound activates TLR3 and the RNA helicase-domain of MDA5, both members of the PAMP family, leading to DC and natural killer (NK) cell activation and production of a "natural mix" of type I interferons, cytokines, and chemokines. Furthermore, poly-ICLC exerts a more direct, broad host-targeted anti-infectious and possibly antitumor effect mediated by the two IFN-inducible nuclear enzyme systems, the 2'5'-OAS and the Pl/eIF2a kinase, also known as the PKR (4-6), as well as RIG-I helicase and MDA5.

**[0172]** In rodents and non-human primates, poly-ICLC was shown to enhance T cell responses to viral antigens, cross-priming, and the induction of tumor-, virus-, and autoantigen-specific CD8+ T-cells. In a recent study in non-human primates, poly-ICLC was found to be

essential for the generation of antibody responses and T-cell immunity to DC targeted or non-targeted HIV Gag p24 protein, emphasizing its effectiveness as a vaccine adjuvant.

**[0173]** In human subjects, transcriptional analysis of serial whole blood samples revealed similar gene expression profiles among the 8 healthy human volunteers receiving one single s.c. administration of poly-ICLC and differential expression of up to 212 genes between these 8 subjects versus 4 subjects receiving placebo. Remarkably, comparison of the poly-ICLC gene expression data to previous data from volunteers immunized with the highly effective yellow fever vaccine YF17D showed that a large number of transcriptional and signal transduction canonical pathways, including those of the innate immune system, were similarly upregulated at peak time points.

**[0174]** More recently, an immunologic analysis was reported on patients with ovarian, fallopian tube, and primary peritoneal cancer in second or third complete clinical remission who were treated on a phase 1 study of subcutaneous vaccination with synthetic overlapping long peptides (OLP) from the cancer testis antigen NY-ESO-1 alone or with Montanide-ISA-51, or with 1.4 mg poly-ICLC and Montanide. The generation of NY-ESO-1 -specific CD4+ and CD8+ T-cell and antibody responses were markedly enhanced with the addition of poly-ICLC and Montanide compared to OLP alone or OLP and Montanide.

**[0175]** Lot release testing for the batches of DCs may be performed to meet minimum specifications before the DCs are injected into patients (see e.g., Sabado et al. (2013) Preparation of tumor antigen-loaded mature dendritic cells for immunotherapy, J. Vis Exp. Aug 1;(78). doi: 10.3791/50085).

**Indications**

**[0176]** In one embodiment, the subject is suffering from a neoplasia selected from the group consisting of: Non-Hodgkin's Lymphoma (NHL), clear cell Renal Cell Carcinoma (ccRCC), melanoma, sarcoma, leukemia or a cancer of the bladder, colon, brain, breast, head and neck, endometrium, lung, ovary, pancreas and prostate.

**[0177]** In one embodiment, the subject is suffering from an autoimmune disease selected from the group consisting of: celiac disease, diabetes mellitus type 1, Graves disease, inflammatory bowel disease, multiple sclerosis, psoriasis, rheumatoid arthritis, and systemic lupus erythematosus.

**[0178]** In one embodiment, the subject is suffering from a bacterial disease or is at risk of contracting a bacterial disease selected from the group consisting of the genus's: Bacillus, Bartonella, Bordetella, Borrelia, Brucella, Campylobacter, Chlamydia and Chlamydophila, Clostridium, Corynebacterium, Enterococcus, Escherichia, Francisella, Haemophilus,

Helicobacter, Legionella, Leptospira, Listeria, Mycobacterium, Mycoplasma, Neisseria, Pseudomonas, Rickettsia, Salmonella, Shigella, Staphylococcus, Streptococcus, Treponema, Ureaplasma, Vibrio, and Yersinia.

**[0179]** In one embodiment, the subject is suffering from a viral disease or is at risk of contracting a viral disease selected from the group consisting of the families: Adenoviridae, Picornaviridae, Herpesviridae, Hepadnaviridae, Flaviviridae, Retroviridae, Orthomyxoviridae, Paramyxoviridae, Papovaviridae, Polyomavirus, Rhabdoviridae, and Togaviridae.

**[0180]** In one embodiment, the subject is suffering from a protazoan disease or is at risk of contracting a protazoan disease, such as, but not limited to malaria.

**Epitope Prediction Coupled with TCR Profiling**

**[0181]** In certain example embodiments, the epitope identification methods disclosed herein may be coupled with TCR profiling to derive antigen-specific therapeutics tailored to the immunodominance of a particular epitope. For example, the methods disclosed herein may be used to derive protective vaccines as well as tolerizing vaccines that induce or more limited immune response than a protective vaccine. A sample to be analyzed is obtained. The sample may originate from a patient to be treated or may be representative of a condition to be treated. The sample may be derived from a biological sample or a biopsy sample depending on the nature of the disease to be addressed. The sample may be an infectious disease sample, a cancer tumor sample, or may be derived from inflamed tissues associated with inflammatory disease, autoimmunity, or allergies. A portion of the sample may be used to derive an input genomic sequence for epitope identification in accordance with the methods described above. In addition, the sample may be used for tissue profiling analysis, such as tissue transcriptome, metatranscriptome, or tumor transcriptome analysis to identify the cell types and/or cell states present in the sample.

**[0182]** Transcriptome analysis may be earned out using known methods in the art. For example, the transcriptome analysis may involve high-throughput single-cell RNA-seq and/or targeted nucleic acid profiling (for example, sequencing, quantitative reverse transcription polymerase chain reaction, and the like).

**[0183]** In certain embodiments, the invention involves single cell RNA sequencing (see, e.g., Kalisky, T., Blainey, P. & Quake, S. R. Genomic Analysis at the Single-Cell Level. Annual review of genetics 45, 431-445, (201 1); Kalisky, T. & Quake, S. R. Single-cell genomics. Nature Methods 8, 311-314 (201 1); Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research, (201 1); Tang, F. et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nature

Protocols 5, 516-535, (2010); Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6, 377-382, (2009); Ramskold, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature Biotechnology 30, 777-782, (2012); and Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports, Volume 2, Issue 3, p666-673, 2012).

**[0184]** In certain embodiments, the invention involves plate based single cell RNA sequencing (see, e.g., Picelli, S. et al., 2014, "Full-length RNA-seq from single cells using Smart-seq2" Nature protocols 9, 171-181, doi:10.1038/nprot.2014.006).

**[0185]** In certain embodiments, the invention involves high-throughput single-cell RNA-seq and/or targeted nucleic acid profiling (for example, sequencing, quantitative reverse transcription polymerase chain reaction, and the like) where the RNAs from different cells are tagged individually, allowing a single library to be created while retaining the cell identity of each read. In this regard reference is made to Macosko et al., 2015, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets" Cell 161, 1202-1214; International patent application number PCT/US20 15/049178, published as WO2016/040476 on March 17, 2016; Klein et al., 2015, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells" Cell 161, 1187-1201; International patent application number PCT/US20 16/027734, published as WO2016168584A1 on October 20, 2016; Zheng, et al., 2016, "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing" Nature Biotechnology 34, 303-31 1; Zheng, et al., 2017, "Massively parallel digital transcriptional profiling of single cells" Nat. Commun. 8, 14049 doi: 10.1038/ncommsl4049; International patent publication number WO2014210353A2; Zilionis, et al., 2017, "Single-cell barcoding and sequencing using droplet microfluidics" Nat Protoc. Jan;12(l):44-73; Cao et al., 2017, "Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing" bioRxiv preprint first posted online Feb. 2, 2017, doi: dx.doi.org/10. 1101/104844; Rosenberg et al., 2017, "Scaling single cell transcriptomics through split pool barcoding" bioRxiv preprint first posted online Feb. 2, 2017, doi: dx.doi.org/10. 1101/105163; Vitak, et al., "Sequencing thousands of single-cell genomes with combinatorial indexing" Nature Methods, 14(3):302-308, 2017; Cao, et al., Comprehensive single-cell transcriptional profiling of a multicellular organism. Science, 357(6352):661-667, 2017; and Gierahn et al., "Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput" Nature Methods 14, 395-398 (2017), all the contents and disclosure of each of which are herein incorporated by reference in their entirety.

**[0186]** In certain embodiments, the invention involves single nucleus RNA sequencing. In this regard reference is made to Swiech et al., 2014, "In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9" Nature Biotechnology Vol. 33, pp. 102-106; Habib et al., 2016, "Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons" Science, Vol. 353, Issue 6302, pp. 925-928; Habib et al., 2017, "Massively parallel single-nucleus RNA-seq with DroNc-seq" Nat Methods. 2017 Oct;14(10):955-958; and International patent application number PCT/US2016/059239, published as WO2017164936 on September 28, 2017, which are herein incorporated by reference in their entirety.

**[0187]** Any method of HLA typing known in the art may be used for the present invention (see, e.g., Erlich, HLA typing using next generation sequencing: An overview, Human Immunology 76 (2015) 887-890).

**[0188]** A second portion from the same sample may be used for T-cell profiling. In certain embodiments, the invention involves identifying T cell receptors (TCR) specific for activation by an immunodominant epitope of the present invention. Identifying, cloning and expressing TCRs according to the present invention may be performed according to any methods known in the art. CD4 + T cells may be isolated by sorting for CD4+ cells. Regarding TCR sequencing, single cell TCR sequencing and expression of TCRs on an immune cell, reference is made to International patent application, "Methods of isolating T cell receptors," serial number PCT/US20 15/067 154 and published as WO2016100977A1; Rosati et al., Overview of methodologies for T-cell receptor repertoire analysis BMC Biotechnol. 2017, 17:61; and Calis and Rosenberg, Characterizing immune repertoires by high throughput sequencing: strategies and applications Trends Immunol. 35(12): 581-590, incorporated herein in its entirety. T cells may be profiled, for example using single cell RNA-seq to differentiate between T cells of a pathogenic and protective phenotype. The sample may be further analyzed using single cell TCR-seq to identify pathogenic TCRs and protective TCRs. Single cell RNA-seq targeted for TCRs (TCRseq) is described further in materials and methods.

**[0189]** Epitopes identified using the methods disclosed herein may be cloned into a first peptide library, such as a lentiviral peptide MHOIIζζ library using known methods in the art. Likewise, the identified TCRs may be cloned into a second library, such as a lentiviral TCRζζ library. The first and second libraries may then be transfected into appropriate cell types to allow expression of the peptide-MHCII and TCR complexes on localization on the cell membrane respectively. For example, the peptide-MHCII library may be clone into a BW5147 cell line, and the TCR library may be cloned into a BW5 147-CD4.2 cell line. One cell line may

further be modified to include a reporter system that generates a detectable signal upon binding of the TCR and peptide-MHC complexes. For example, the cell line comprising the MHCII library may further include a reporter construct. In certain example embodiments, the reporter construct may comprise a GFP under control of an inducible promoter. The inducible promoter may be under the control of or responsive to a signal generated by an activated TCR or MHCII complex. For example, the reporter construct may the inducible promoter may activate expression of GFP in response to IL-2 production. While GFP is provided by way of example, other detectable signals that facilitate sorting and or isolation of the peptide-MHCII expressing cells may also be used. The TCR expressing and peptide-MHCII expressing cells are brought into contact with one another under conditions to allow binding between complementary TCR and MHCII complexes. Activated MHCII expressing cells are then sort or otherwise isolated based on expression of the detectable signal. The MHC peptides are subsequently isolated and sequenced using known methods in the art. The number of peptide sequencing reads per peptide provides a measure of each epitopes immunodominance. The present invention may use any known method of peptide sequencing known in the art (see, e.g., U.S. Patent number 6,379,970).

**[0190]** In certain example embodiments, peptides with the requisite or desired degree of immunodominance are then isolated and further used to prepare immunogenic compositions as described above. Said peptides may also be used for Ag-specific expansion of endogenous T-cells and/or Ag-specific depletion of endogenous T cells using known methods in the art.

**Probing the Host-Commensal Relationship to Reveal "Health Status" of the Immune System**

**[0191]** The human microbiota is the aggregate of microorganisms that resides on or within any of a number of human tissues and biofluids, including the skin, mammary glands, placenta, seminal fluid, uterus, ovarian follicles, lung, saliva, oral mucosa, conjunctiva, biliary and gastrointestinal tracts. They include bacteria, archaea, fungi, protists and viruses. Some microorganisms that colonize humans are commensal, meaning they co-exist without harming humans.

**[0192]** In certain embodiments, monitoring the magnitude and nature of the T cell response to specific commensal antigens can reveal the health status of the immune system. For example, having Tregs reactive to commensal epitopes may indicate health, having Thl/Thl7 cells reactive to commensal epitopes may indicate autoimmunity, having Tregs reactive to tumor antigen epitopes may indicate cancer status (e.g., a suppressed immune response to the tumor), and having Th2 cells reactive to allergen epitopes may indicate allergy (see, **Fig. 20).** In certain

embodiments, pMHCII tetramers are used to track commensal T cells in a subject (see, **Fig. 21).**

**[0193]** In the state of health, the adaptive immune system maintains tolerance to local gut antigen exposure and protection from systemic infection. The sheer number of bacterial species and diversity of protein-coding elements in the microbiome represents an enormous search space. Towards identifying immunogenic commensals, Applicants developed "serum immunoglobulin commensal capture and sequencing" (SICC-seq). In certain embodiments, commensal epitopes are predicted from sequencing data obtained from a biological sample. For example, a sample may be taken from a subject (e.g., stool) and the microbiota may be identified. Genomes of organisms in the microbiota may be identified and MHCII epitopes predicted according to the present invention (e.g., according to HLA-type of the subject). In certain embodiments, the HLA type of the subject is determined using a blood sample (e.g., serum). In certain embodiments, the epitopes are used to detect cytokine production in immune cells (e.g., T cells) obtained from the subject (e.g., IL-2, IFN-γ, IL-17, IL-10) (see, **Fig. 21).** Cytokine assays (e.g., cytokine arrays) are well known in the art. In certain embodiments, IL-10 secretion in response to an epitope indicates health (e.g., secretion from Tregs). In certain embodiments, secretion of IL-17 (e.g., Thl7 cells) indicates autoimmunity. In certain embodiments, pMHCII tetramers are used to track commensal T cells secreting the cytokines in the subject (see, **Fig. 21).**

**[0194]** The present invention has advantageously used MHCII peptidomics to identify novel bacterial antigens. Additionally, Applicants have determined novel features of antigenicity from immunopeptidome generally applicable to the prediction of any MHCII binding peptide. Moreover, Applicants have advantageously provided a deep neural network algorithm, BOTA, that predicts immunodominant epitopes. Finally, Applicants have determined that host autophagy pathways differentially control epitope selection.

**[0195]** Although specific embodiments have been described above in detail, the description is merely for purposes of illustration. It should be appreciated, therefore, that many aspects described above are not intended as required or essential elements unless explicitly stated otherwise. Modifications of, and equivalent components or acts corresponding to, the disclosed aspects of the example embodiments, in addition to those described above, can be made by a person of ordinary skill in the art, having the benefit of the present disclosure, without departing from the spirit and scope of embodiments defined in the following claims,

the scope of which is to be accorded the broadest interpretation so as to encompass such modifications and equivalent structures.

**[0196]** This invention is further illustrated by the following examples, which are not to be construed in any way as imposing limitations upon the scope thereof. On the contrary, it is to be clearly understood that resort may be had to various other embodiments, modifications, and equivalents thereof which, after reading the description herein, may suggest themselves to those skilled in the art without departing from the spirit of the present invention and/or the scope of the appended claims.

## EXAMPLES

### Example 1 - MHCII peptidomics in primary murine dendritic cells defines biochemical features of antigenicity

**[0197]** Defining the immunodominance hierarchy of T cell epitopes remains a significant challenge in the context of infectious disease, autoimmunity, and immune oncology. Even for some of the most widespread bacterial pathogens, little is known about which antigens drive protective CD4 T cell responses. Towards the aim of defining features of antigenicity, Applicants developed a proteomics-based approach to identify MHCII-associated peptides in murine bone marrow-derived dendritic cells (BMDCs) from C57BL/6J mice expressing I-$A^b$. In these experiments, mature peptide-MHCII complexes were immunoprecipitated from primary BMDCs with monoclonal Y3P antibody (Janeway et al., 1984). Associated peptides from biological replicates were acid-eluted, chemically labeled with iTRAQ isobaric mass tags for relative quantification, mixed, and analyzed using high resolution, accurate mass liquid chromatography-tandem mass spectrometry (LC-MS/MS) **(Fig. 7A).** This approach identified 3,671 unique peptides derived from 1,088 autologous murine proteins **(Fig. 7B** and SEQ ID NOs: 1-14979). The sensitivity afforded by Orbitrap detection methods in combination with differential mass tagging by iTRAQ enabled relative quantitation of significantly more peptides than previously reported approaches.

**[0198]** Having generated an expansive catalogue of the MHCII immunopeptidome in murine BMDCs, Applicants sought to identify key biochemical features of antigenicity. Initially, Applicants derived the optimal I-$A^b$-binding motif from primary sequence features in autologous murine peptides **(Fig. 7C).** The resulting 9mer core peptide sequence resembles previous predictions based on *in vitro* binding kinetics between synthetic peptide libraries and purified I-$A^b$ (Andreatta et al., 201 1). In contrast, Applicants detected a strong preference for proline in the P4 position.

### Example 2 - Autophagy shapes the MHCII immunopeptidome

**[0199]** The primary sequence of antigenic peptides confers binding to MHCII but does not represent the only factor governing antigenicity. The nature of how host pathways shape the immunopeptidome remains incompletely understood. Given that lysosomal pathways contribute to antigen processing and epitope selection, Applicants hypothesized that autophagy shapes the immunopeptidome by dictating which antigens gain access to the lysosomal compartment. Applicants aimed to shed light on the subject by leveraging MHCII peptidomics. The core autophagy protein ATG16L1 is required for macroautophagy and xenophagy, which divert cytosolic cargo to lysosomes for disposal. To selectively perturb autophagy in antigen-presenting cells, Applicants generated *Atgl6ll* $^{f/f}$ x CDl lc-Cre mice *(Atgl6ir'-).* MHCII peptidomics experiments demonstrated that *Atg161l-r* BMDCs were severely impaired in their ability to present peptides derived from organelle-derived and cytosolic proteins **(Fig. 8A and 8B).** In contrast, *Atgl6U*$^{\sim}$cells displayed elevated levels of lysosome- and endosome-derived peptides compared to WT BMDCs **(Fig. 8B and 8E).** These findings are consistent with the known role for Atg 1611 in macroautophagy and highlight the degree to which lysosomal trafficking impacts the spectrum of peptides presented by MHCII.

**[0200]** Given the abundance of MHCII-associated peptides derived from endolysosomal proteins, Applicants queried these sequences to define additional features of antigenicity. Using the high-resolution I-A$^{b}$-binding motif as a reference sequence **(Fig. 7C),** Applicants scanned endolysosomal proteins identified by proteomics for consensus. As expected, Applicants were able to retrospectively predict peptides that were identified by mass spectrometry; however, Applicants found many instances of consensus sites that were not presented on MHCII. In the vast majority of these cases, consensus sites located within structured protein domains were not detected by proteomics, whereas those located in unstructured inter-domain regions were. For example, MHCII peptidomics identified one peptide derived from the murine cation-independent mannose 6-phosphate receptor (CEVIR). This peptide is located between two CEVIR domains. In contrast, Applicants found three additional peptides conforming to the I-A$^{b}$-binding consensus motif that were located within CIMR domains and were not detected as processed peptides bound to MHCII **(Fig. 8C).** Across the entire dataset, epitopes preferentially derived from interdomain regions of greater than 20 amino acids or protein domains greater than 30 amino acids **(Fig. 13 and 16).** Thus, epitope accessibility in the context of native proteins is an important factor that impacts MHCII binding and protease processing.

**Example 3 - Training a deep neural network on MHCII peptidomics data**

**[0201]** Having demonstrated that MHCII peptidomics can identify antigenic features conferred by lysosomal processing, Applicants leveraged this dataset to devise a neural network-based algorithm for epitope prediction. Historically, predicting MHCII binding strength for a given peptide has been challenging. The concept of epitope accessibility, together with MHCII affinity, adds to the complexity of formulating predictions for immunodominant epitopes. Such features are not adequately captured by linear relationships; rather, a non-linear architecture is more appropriate. Therefore, Applicants explored the use of a deep neural network-based algorithm to predict these epitopes using peptidomics data as a training set.

**[0202]** Applicants developed a model, bacteria-origin T cell antigen (BOTA) predictor, that generates a list of candidate peptides utilizing information including protein cellular location, transmembrane structure (if applicable), and domain distribution **(Fig. 8D).** BOTA requires only an annotated genome input to extract amino acid sequences of protein-coding genes. BOTA then relies on outputs from three algorithms: HMMTOP, PSORT, and HMMER's search against pfam. First, BOTA identifies secreted and cell wall proteins (Yu et al., 2010). Second, it masks intracellular regions and transmembrane domains of cell wall proteins (with an 8-amino acid flanking buffer) (Tusnady and Simon, 1998). Third, it excludes regions that fall within small domains or between a series of adjacent domains (inaccessible, compact folding) (Marchler-Bauer et al., 2015). The metric used for domain mapping includes the distance to the up/downstream domains, the density of flanking domains, and the domain size. Finally, candidate protein regions are piped into the deep neural network predictor to generate their probabilities to be MHCII binders **(Fig. 8D).** Applicants note that this pipeline is modular and can be used to generate allele-specific models in both human and model animals.

**[0203]** The core algorithm of BOTA is a deep neural network, which employs a sparse representation of input peptides and multiple pre-trained binder models to generate a feature map. The feature map includes many nonlinear summarizations of peptide features that are used to make predictions. The output from the deep neural network is a score (f) that serves as an indicator of MHCII binding. In the case of mouse I-$A^b$, Applicants used proteomics data to train the BOTA model. The model was trained using randomly sampled parameters 30 times on a 3-fold cross-validation scheme. The optimal parameter calibration was selected based on their receiver operating characteristic curves (ROC AUCs).

**[0204]** The advantages of BOTA in comparison to current methods are that BOTA (1) requires only whole genome, not polypeptides, as input, (2) considers epitope accessibility by focusing the search on extracellular regions of proteins and epitope location relative to protein domain organization, (3) is trained on peptidomics data from antigen-presenting cells, which

are significantly more accurate than *in vitro* peptide binding data, and (4) easily scales to thousands of genomes for different alleles. Thus, BOTA is capable of accurately predicting immunogenic epitopes from medically relevant bacterial pathogens.

**Example 4 - Validation of BOTA epitope predictions for *Listeria* with MHCII peptidomics**

**[0205]** To validate BOTA, Applicants tested its performance in predicting MHCII-restricted epitopes from *Listeria monocytogenes*, one of the most common foodborne pathogens associated with a relatively high mortality rate (Scallan et al., 2011). Despite the epidemiological impact of listeriosis, the number of experimentally validated CD4 T cell epitopes encoded within the *Listeria* genome is limited. Thus, Applicants first sought to identify MHCII-associated peptides in BMDCs exposed to live *Listeria* by peptidomics. In these experiments, primary BMDCs were exposed to live *L. monocytogenes* for 10 minutes or 6 hours prior to immunoprecipitation of mature peptide-MHCII complexes with monoclonal Y3P antibody (Janeway et al., 1984). Associated peptides from biological replicates of the 10-minute and 6-hour timepoints were analyzed using mass spectrometry (LC-MS/MS) **(Fig. 9A and B).** These experiments discovered 48 unique peptides derived from exogenous *Listeria* proteins. Twenty-nine of these peptides represented nested sets derived from 7 unique proteins **(Table 1).** The second-most enriched peptide was the previously identified immunodominant epitope from listeriolysin O ($LLO_{190-205}$), while the remaining peptides represented novel candidate antigens. Notably, all of the detected peptides were derived from secreted proteins or cell wall proteins. These observations highlight the importance of epitope accessibility for antigen presentation. Together, this dataset represents the first example of identifying processed peptide antigens derived from a live bacterial pathogen.

**[0206]** Applicants next compared *Listeria* peptides identified by MHCII peptidomics with BOTA predictions. Nine out of 17 BOTA-predicted epitopes were validated by MHCII peptidomics **(Table 2).** Of the 35 unique *Listeria* peptides identified by proteomics with an adjusted p value<0.05, BOTA-predicted epitopes were present in 28. Applicants compared the prediction accuracy of the BOTA model between training with two data sets: peptides captured in murine dendritic cells ("Peptidomics Training") and MHCII-associated peptides from the Immune Epitope Database ("IEDB Training") **(Fig. 9D).** Both BOTA pre-training accuracies plateaued after 200 epochs, but BOTA training with the peptidomics data increased prediction accuracy over training with IEDB data by 15%. Notably, BOTA showed strong improvement over the current prediction methods, including the state-of-the-art netMHCIIpan (Andreatta et

al., 201 1) **(Fig. 9E).** Thus, BOTA performs well in predicting dominant CD4 T cell epitopes **(Table 2)**.

**[0207]** Improvement in prediction accuracy by BOTA also signifies the successful application of a deep neural network to a complex biomedical problem. Previous efforts using traditional neural networks or hidden Markov models were limited by their ability to extract highly abstract features, thus leading to insufficient insights into epitope prediction (Wang et al., 2008). The important innovation of BOTA is its utility in predicting CD4 T cell epitopes for virtually any MHCII allele and any antigen source, including commensal microbes, pathogens, autoantigens, and tumor antigens.

**Example 5 - BOTA and MHCII peptidomics predict immunodominance of *Listeria* epitopes *in vivo***

**[0208]** Given that MHCII peptidomics and BOTA successfully identified *Listeria* peptides and features associated with efficient antigen presentation, Applicants sought to measure T cell responses to these epitopes *in vivo.* To quantify the CD4 T cell response directed to the top eight candidate *Listeria* epitopes, mice were sacrificed seven days after intraperitoneal (i.p.) infection with *Listeria.* T cells were restimulated *in vitro* with synthetic peptides, and IFN-γ responses were enumerated by ELISPOT. Robust responses were detected against four of the eight epitopes (lmo0202, lmo2558, lmo2185, lmo0135) **(Fig. 10A).** The remaining four epitopes elicited weaker responses near the limit of detection for ELISPOT **(Fig. 10B).** Notably, the T cell response to these eight candidate epitopes correlated remarkably well with their abundances detected by MHCII peptidomics **(Fig. IOC).** RNA expression of *Listeria* antigens (Chatterjee et al., 2006) correlated with the corresponding T cell response to a lesser extent **(Fig. 10D),** although low expression did not appear to preclude antigenicity **(Fig. 10E).**

**Example 6 - Integration of T cell phenotype with TCR specificity in the *Listeria* response**

**[0209]** Having demonstrated the utility of BOTA for predicting bacterial epitopes, Applicants next sought to rigorously characterize the antigen-specific T cell response to *Listeria.* Standard approaches for defining immunodominance, such as ELISPOT, rely on ranking candidate epitopes based on the magnitudes of the T cell responses they elicit *in vivo.* However, such approaches require prior knowledge of TCR specificity and T cell phenotype (cytokine profile) and cannot account for T cell responses to epitopes that have not been tested, but may be dominant. To address these limitations, Applicants developed a procedure for single cell analysis of T cell phenotypes matched to their corresponding TCR sequences. This TCRseq approach allows for simultaneous enumeration of T cell clonal frequency and cytokine profiles **(Fig. 17).** Dominant TCR clones can then be screened for reactivity against *Listeria* epitopes.

**[0210]** Towards this end, Applicants infected mice with *Listeria* and FACS-sorted single CD4 T cells for transcriptomics and TCR sequencing. Based on transcriptional profiles, T cells clustered into distinct functional states **(Fig. 11A),** with a distinct profile for activated Thl cells expressing Ifnγ **(Fig. 11B)** and Tbx21 **(Fig. 11C).** Amongst these T cells, the TCR repertoire was remarkably diverse with a maximal clonal frequency of -2% **(Fig. 11D).**

**[0211]** Based on unbiased clustering of transcriptional profiles, T cells partitioned into four distinct subsets **(Fig. 12A).** To identify the activated T effector (Teff) cluster, Applicants generated a per-cell Teff score that was based on the gene expression signature derived from ImmGen datasets comparing CD4 Teff splenocytes (day 8 after lymphocytic choriomeningitis virus, LCMV, infection) versus naive CD4 T splenocytes (Heng et al., 2008). These analyses identified cluster 2 as being enriched for Teff cells expressing signature activation markers (Ccl5, Nkg7, Ikzfl), genes that regulate ER homeostasis (Calr, Pdia3), and genes that control cellular metabolism (Acly, Akrlal) **(Fig. 12B).** Amongst these T cells, the TCR repertoire was remarkably diverse with a maximal clonal frequency of -3% for conventional T cells. The most abundant conventional TCR (TRAV14-2TRAJ25|TRBV5TRBJ2-7) was detected exclusively in cells residing in Teff cluster 2. By contrast, NKT cells expressing the invariant alpha chain Travl lTraj 18 were found scattered throughout clusters **(Fig. 12C).**

**[0212]** Using TCRseq and whole transcriptome data, Applicants prioritized the most abundant TCR to test for reactivity against BOTA predictions **(Fig. 13A and 14A).** In designing a screening modality, Applicants took into consideration the low affinity interaction between TCRs and peptide antigen presented on MHCII. This interaction occurs between multiple TCRs and pMHCII complexes within lipid bilayers on adjacent cells. Avidity and kinetics of engagement/disengagement elicit a TCR signaling cascade that amplifies the input signal. Therefore, Applicants designed a heterologous expression system in which TCR-negative BW5147-CD4-CD28 cells are transduced to express single chain TCRa and TCRP fused to the cytoplasmic tail of CD3ζ. Functional TCRap proteins in BW5 147-CD4-CD28 cells engage cognate peptide-MHCII on antigen presenting cells to initiate a TCR signaling response that results in expression of IL-2 or the activation marker *Nur77 {Nr4al).* As a source of surrogate antigen presenting cells, Applicants utilized HEK293T cells or BW5147-B7/4 cells transfected or transduced to express I-$A^{b}$α-CD3ζ and candidate peptide antigens fused to I-$A^{b}$β-CD3ζ. For HEK293T experiments, Applicants selected the most abundant TCR clone from TCRseq to test for reactivity with 4 *Listeria* pMHCII complexes predicted by BOTA **(Fig. 14B).** For BW5147-B7/4 experiments, Applicants generated 4 TCRs from TCRseq showing

high IFN-γ expression and 2 *Listeria* pMHCII complexes predicted by BOTA **(Fig. 13B).** As a positive control, Applicants demonstrated that the OT2 TCR expressed in BW5147-CD4-CD28 cells reacted with ovalbumin peptide in the context of I-$A^b$ by inducing IL-2 secretion or *Nur77* expression **(Fig. 13B and 14B)** and IL-2 production. Among the four candidate *Listeria-speciiic* TCRs, three showed evidence of weak reactivity to LLO **(Fig. 13B).** Similarly, a previously identified TCR (LL0 _118) (Weber et al., 2012) reacted robustly with LLO **(Fig. 14B).** The top TCR candidate identified by TCRseq also reacted with LLO, according to induction of IL-2 and binding to LLO-I-$A^b$ tetramers **(Fig. 14B and C).** These findings are consistent with the notion that the primary T cell response to *Listeria* is characterized by numerous weakly-reactive TCRs. In agreement with this conclusion, the TCRseq experiments revealed a remarkably diverse TCR repertoire associated with the primary Thl response to *Listeria.* Taken together, these experiments establish the feasibility of integrating population-level TCR repertoires with antigen specificity by screening individual TCRs for reactivity against candidate antigens predicted *in silico.*

**Example 7 - Identification of commensal epitopes**

**[0213]** Having validated BOTA epitope predictions for a common pathogenic bacterium, Applicants sought to explore the complex relationship between commensal microbes and host adaptive immunity. In this context, the intestinal microbiome fine-tunes inflammatory thresholds, primes innate immune effector function, and shapes the adaptive immune response through selection and tolerization of lymphocytes (Palm et al., 2015)). Thus, the dynamic role of the microbiome in immune education impacts organ systems throughout the body, and in turn, many disease states (Hall et al., 2017). Applicants reasoned that because the host adaptive immune system continuously interacts with the microbiome, monitoring the magnitude and nature of the T cell response to specific commensal antigens can reveal the health status of the immune system. In the state of health, the adaptive immune system maintains tolerance to local gut antigen exposure and protection from systemic infection. The sheer number of bacterial species and diversity of protein-coding elements in the microbiome represents an enormous search space. Towards identifying immunogenic commensals, Applicants developed "serum immunoglobulin commensal capture and sequencing" (SICC-seq). Mice were administered a course of dextran sodium sulfate (DSS) to induce barrier breach and allowed to recover for 7 days. Serum was then harvested from mice and incubated with stool to opsonize commensals. IgG-positive microbes were enriched by selection with Protein A/G-coupled magnetic beads and analyzed along with total stool by 16S rRNA sequencing. Using this approach, Applicants

demonstrated that induction of colitis with DSS elicits a systemic T cell-dependent IgG response that preferentially targets bacteria in the genus Bacteroidales **(Fig. 15A).** In contrast, Akkermansia evade this response, likely due to their ability to induce T cell tolerance and IgA responses under homeostatic conditions (before colitis) **(Fig. 15A).** Having established the immunogenicity of Bacteroidales, Applicants employed BOTA to predict T cell epitopes from the dominant species inhabiting mice (Ormerod, et al., 2016) **(Fig. 15B)** and identified a highly abundant epitope in a SusC-like protein that is conserved across the Bacteroidales genus and often duplicated within species **(Fig. 15B and 22).** To determine if T cells recognize this SusC epitope *in vivo,* Applicants harvested splenocytes from a naive mouse and stimulated them with SusC peptide *in vitro.* Importantly, the SusC peptide induced IL-10 production by T cells, indicating a homeostatic relationship between host T cells and Bacteroidales in a normal healthy mouse **(Fig. 15C).** Applicants hypothesize that these interactions extend to many other commensals, and that tumultuous relationships between T cells and commensals typify immune pathologies and autoimmunity.

**Example 8 - Discussion**

**[0214]** Host-microbe interactions cooperatively influence the specificity and diversity of the T cell response. A deeper understanding of this relationship requires new approaches for unbiased antigen discovery, defining the features of antigenicity, and elucidating host pathways underlying preferential selection of these features. Toward these objectives, Applicants report a highly quantitative adaptation of MHCII peptidomics. By recapitulating the interaction between primary murine BMDCs and live *L. monocytogenes,* Applicants identified four dominant CD4 T cell epitopes and established their immunodominance hierarchy *in vivo.* In addition, MHCII peptidomics identified over 3600 autologous mouse peptides, which provided an exceptionally detailed view of lysosomal function.

**[0215]** As a consequence of deep profiling of the MHCII immunopeptidome, Applicants generated a rich dataset for identifying features associated with of antigen processing and developed BOTA as a predictive algorithm, incorporating several important attributes of immunodominant epitopes revealed by proteomics. Based on observations from MHCII peptidomics, optimal epitopes tend to (1) derive from secreted proteins or extracellular regions of cell wall proteins in bacteria, (2) have a primary sequence structure that conforms to a defined MHCII binding motif, (3) be located more than 8 amino acids away from transmembrane domains, and (4) have tertiary structure characteristics that promote accessibility to the MHCII binding groove and the enzymes required for proteolytic processing. Notably, Applicants did not identify any features of antigenicity associated with protease

specificity, as lysosomal endopeptidases and exopeptidases are highly diverse with respect to substrate selectivity. Importantly, all of the features identified by peptidomics are readily identifiable at the level of DNA sequence; therefore, the only input requirement for BOTA is an annotated genome. Owing to this, BOTA enables large-scale mapping of the immunodominant epitope landscape for any bacterial species or collection of species, such as the human microbiome. Taken together, Applicants demonstrate the utility of MHCII peptidomics for training a deep neural network that specifically identifies candidate epitopes with key features associated with immunodominance and antigenicity. Furthermore, MHCII peptidomics serves as a powerful tool for unbiased discovery of complex pathogen antigens and for interrogation of host pathways underlying human disease.

**[0216]** Inherent to antigen discovery is the significant challenge of validating epitope predictions. Conclusive validation of epitope immunogenicity requires demonstration that a measurable T cell response is elicited *in vivo.* To address this challenge, Applicants developed an approach for identifying and integrating TCR repertoire, phenotype, and antigen reactivity. Coupling TCRseq with whole transcriptome profiling enabled at the single cell level enabled assignment of transcriptional phenotypes to individual TCRs. While Applicants employed this approach to identify TCRs associated with CD4 T effector cells (e.g., Thl cells) derived from *Listeria*-infected mice, it is applicable to any immune cell type and any phenotype that can be defined at the level of the transcriptome, including Treg or Thl7. Such determinations are challenging in T cell hybridomas because the primary T cell phenotype is not preserved after immortalization. Moreover, generating T cell hybridomas is an inefficient process that is further biased by chromosome loss, drug selection, and screening for antigen reactivity. In contrast, TCRseq is comparatively efficient, which is an important consideration in the context of limited T cell input from precious clinical specimens. A small tissue biopsy is sufficient to generate a permanent archive of TCR sequences matched with transcriptional profiles that can be used to screen defined TCRs for reactivity against candidate epitopes. Thus, immunodominance can be unambiguously defined in the context of epitopes that elicit the strongest T cell responses, as determined by clonal frequency and/or absolute numbers.

**[0217]** Here, Applicants develop a technology platform with broad utility for antigen discovery. BOTA is capable of predicting CD4 T cell epitopes from multiple sources, including genomes or transcriptomes derived from pathogens, the microbiome, allergens, tumors, and tissue biopsies **(Fig. 18).** By coupling BOTA predictions with TCRseq from matching specimens, Applicants enable generation of DNA-encoded epitope and TCR libraries that can be screened to define antigen specificity and quantify immunodominance. The screening

platform developed can be implemented in arrayed format (well-based screening of individual TCRs by peptide-MHCII antigens) or in a pooled format where cells expressing TCR libraries are cocultured with cells expressing peptide-MHCII libraries. For example, Applicants have performed TCR fingerprinting by coculturing stimulator cells expressing the OT2 TCR with responder cells displaying a library of ovalbumin peptide mutants expressed as a fusion protein with MHCII ectodomain linked to the CD3ζ cytoplasmic domain. After FACS-sorting activated responder cells (41bb+) that had engaged in a productive interaction with stimulator cells, Applicants sequenced the Ova-MHCI^ libraries to recover the known cognate peptide ligand for the OT2 TCR **(Fig. 19).** In this context, it is possible to simultaneously screen thousands of epitopes for reactivity with a given TCR.

**[0218]** With the new approaches to effectively predict T cell epitopes and validate TCR reactivity as disclosed herein, it is possible to identify antigenic determinants in bacterial pathogens and even complex communities such as the intestinal microbiome. Applicants identified an I-$A^b$-restricted epitope in a SusC-like protein derived from *Bacteroidales* that is associated with IL-10 producing T cells derived from mouse spleen. Notably, these experiments identified systemically circulating T cells with reactivity to a benign commensal species. The findings highlight the intimate relationship between commensals and the host adaptive immune system. In this context, previous studies in mouse models have demonstrated that a systemic T cell-dependent IgG response to commensals is cross-protective against pathogen infection (Zeng et al., 2016). Similarly, IgG-reactivity with commensals has been shown to be widespread in healthy humans and qualitatively perturbed in the context of autoimmunity (Christmann et al., 2015; and Stoll et al., 2014). While the understanding of host-commensal mutualism is in its infancy, future research stands to reveal how this relationship promotes homeostatic immune regulation or precipitates immune dysfunction. New approaches to antigen discovery portend opportunities for biomedical research, including vaccine development (prophylactic or tolerizing) and antigen-specific T cell therapies aimed at expanding or deleting endogenous T cells with defined antigen specificities.

**[0219]** The improvement on prediction accuracy by BOTA also signifies the successful application of deep neural network in solving complex biomedical problems. Previous efforts using traditional neural networks or hidden Markov Models were limited by their ability to extract highly abstract features, thus leading to insufficient insights into epitope prediction. The important innovation of BOTA, is that it can be utilized to predict CD4 T cell epitopes for virtually any MHCII allele and any antigen-source, including commensal microbes, pathogens, autoantigens, and tumor antigens. Taken together, Applicants demonstrate the utility of MHCII

peptidomics for training a deep neural network that specifically identifies candidate epitopes with key features associated with immunodominance. Furthermore, MHCII peptidomics serves as a powerful tool for unbiased discovery of complex pathogen antigens and for interrogation of host pathways underlying human disease.

**Table 1**: *Listeria* epitopes discovered by MHCII peptidomics.

| SEQ ID NO: | Listeria peptide[a] | logFC[b] | adj.P.Val[c] | entry_name | core | scored |
|---|---|---|---|---|---|---|
| 14980 | .....WN**EKYAQAYPN**VS..... | 1.37728 | 7.0568E-18 | lmo0202 (hly) (LLO) | EKYAQAYPN | 9.728E-11 |
| 14981 | .....WN**EKYAQAYPN**VSAKI..... | 1.20234 | 1.0840E-13 | lmo0202 (hly) (LLO) | EKYAQAYPN | 9.728E-11 |
| 14982 | .....VERWN**EKYAQAYPN**VS..... | 0.28494 | 1.6720E-01 | lmo0202 (hly) (LLO) | EKYAQAYPN | 9.728E-11 |
| 14983 | .....ERWN**EKYAQAYPN**VS..... | 0.20689 | 3.6222E-01 | lmo0202 (hly) (LLO) | EKYAQAYPN | 9.728E-11 |
| 14984 | ....APGQETQHY**YGLPVADSA**IDR..... | 0.85608 | 3.3989E-07 | lmo2558 (ami) | YGLPVADSA | 6.385E-11 |
| 14985 | ......VWTKPNK**IEGAQKISA**LSTY..... | -0.26248 | 2.1661E-01 | lmo2558 (ami) | WTKPNKIEG | 1.641E-11 |
| 14986 | .....ADFRYV**FDTAKATAA**SSYPG..... | 1.56572 | 6.8456E-23 | lmo2185 | FDTAKATAA | 1.932E-10 |
| 14987 | .....ADFRYV**FDTAKATAA**SSYPGSDETPP..... | 1.55744 | 1.0148E-22 | lmo2185 | FDTAKATAA | 1.932E-10 |
| 14988 | .....DFRYV**FDTAKATAA**SSYPGSDETPP..... | 0.67777 | 9.6362E-05 | lmo2185 | FDTAKATAA | 1.932E-10 |
| 14989 | .....RYV**FDTAKATAA**SSYPGSDETPPVVNPGETNPP | 0.63438 | 3.1562E-04 | lmo2185 | ETPPVVNPG | 2.208E-10 |
| 14990 | .....YV**FDTAKATAA**SSYPGSDETPP..... | 0.61121 | 5.4781E-04 | lmo2185 | FDTAKATAA | 1.932E-10 |
| 14991 | .....RYV**FDTAKATAA**SSYPGSDETPP..... | 0.27084 | 1.9678E-01 | lmo2185 | FDTAKATAA | 1.932E-10 |
| 14992 | .....YPGSD**ETPPVVNPG**E..... | 0.78521 | 3.8293E-06 | lmo2185 | ETPPVVNPG | 2.208E-10 |
| 14993 | ......SSYPGSD**ETPPVVNPG**ETNPP..... | 0.29186 | 1.5520E-01 | lmo2185 | ETPPVVNPG | 2.208E-10 |
| 14994 | .....YPGSD**ETPPVVNPG**ETN..... | 1.86133 | 3.1731E-32 | lmo2185 | ETPPVVNPG | 2.208E-10 |
| 14995 | ......EVEDLNQ**PLAAHVNYE**..... | 0.15261 | 5.2950E-01 | lmo2185 | PLAAHVNYE | 4.828E-11 |
| 14996 | .....AVDDTTVK**FTLPTVAPA**FENT..... | 0.87086 | 1.9984E-07 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 14997 | ......VDDTTVK**FTLPTVAPA**FENT..... | 0.84368 | 5.3557E-07 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 14998 | .....DDTTVK**FTLPTVAPA**FENT..... | 0.78647 | 3.6883E-06 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 14999 | .....AVDDTTVK**FTLPTVAPA**FENTIK..... | 0.46417 | 1.2352E-02 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 15000 | ......VDDTTVK**FTLPTVAPA**FE..... | 0.45544 | 1.4456E-02 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 15001 | .....DDTTVK**FTLPTVAPA**FE..... | 0.38746 | 4.5231E-02 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 15002 | .....DDTTVK**FTLPTVAPA**FEN..... | 0.25955 | 2.2271E-01 | lmo0135 | FTLPTVAPA | 5.607E-09 |
| 15003 | .....TTVK**FTLPTVAPA**FENT..... | 0.0598 | 8.4126E-01 | lmo0135 | FTLPTVAPA | 5.607E-09 |

| SEQ ID NO: | Listeria peptide[a] | logFC[b] | adj.P.Val[c] | entry_name | core | scored |
|---|---|---|---|---|---|---|
| 15004 | .....DGLT**WHDGKPLTA**DDV..... | 1.01543 | 6.5055E-10 | lmo0135 | WHDGKPLTA | 1.053E-10 |
| 15005 | .....GTKEKV**VATPVSNVS**TSSA..... | 0.4888 | 7.7931E-03 | lmo0186 | VATPVSNVS | 4.923E-10 |
| 15006 | .....TKEKV**VATPVSNVS**TSS..... | -0.2861 | 1.6535E-01 | lmo0186 | VATPVSNVS | 4.923E-10 |
| 15007 | .....KEKV**VATPVSNVS**TSS..... | -0.20956 | 3.5407E-01 | lmo0186 | VATPVSNVS | 4.923E-10 |
| 15008 | .....GTKEKV**VATPVSNVS**T..... | 0.03708 | 9.0585E-01 | lmo0186 | VATPVSNVS | 4.923E-10 |
| 15009 | .....GGINQA**YTGSTALSD**GLN..... | 0.55469 | 2.0548E-03 | lmo2360 | YTGSTALSD | 7.495E-11 |
| 15010 | .....GINQA**YTGSTALSD**G..... | 0.26067 | 2.2070E-01 | lmo2360 | YTGSTALSD | 7.495E-11 |
| 15011 | .....GGINQA**YTGSTALSD**G..... | 0.23863 | 2.7631E-01 | lmo2360 | YTGSTALSD | 7.495E-11 |
| 15012 | .....STVV**VEAGDTLWG**IAQSKGTT..... | 0.48949 | 7.6998E-03 | lmo0582 (iap) | VEAGDTLWG | 3.576E-12 |
| 15013 | .....IAVTA**FAAPTIASA**..... | 0.18053 | 4.4262E-01 | lmo0582 (iap) | FAAPTIASA | 8.274E-09 |
| 15014 | .....VETKA**DIDALSLLS**DKIFVTNQT..... | 1.05574 | 1.1992E-10 | lmo1451 (ispH) | DIDALSLLS | 1.864E-11 |
| 15015 | .....EAKELV**DNAPKALKE**GIA..... | 0.9296 | 2.0391E-08 | lmo0251 (rplL) | DNAPKALKE | 1.319E-10 |
| 15016 | .....SNVNHVLD**IGYGGGSnV**KN..... | 0.227 | 3.0478E-01 | lmo1715 | IGYGGGSnV | 1.796E-12 |
| 15017 | .....**EVVYHNMKQ**DLE..... | -0.17079 | 4.7439E-01 | lmo2411 | EVVYHNMKQ | 5.047E-14 |
| 15018 | DGLKAGEHFFAAG**KDWPDAIYA**NGDEVAAGVmYH..... | -0.42712 | 2.3527E-02 | lmo2737 | KDWPDAIYA | 4.863E-11 |
| 15019 | .....E**ILLQKPnGE**RLTSFITL..... | -0.04842 | 8.7532E-01 | lmo0429 | ILLQKPnGE | 2.664E-12 |
| 15020 | .....GLLm**AKKYKMNLT**IVIVN..... | -0.62908 | 3.5620E-04 | lmo1675 (menD) | AKKYKMNLT | 1.150E-12 |
| 15021 | .....TVRAKS**KTYPVYInE**FALE..... | 0.14445 | 5.5769E-01 | lmo1927 (aroB) | KTYPVYInE | 1.187E-10 |
| 15022 | .....mNR**VmLPLQQmG**AKMHGK..... | 0.14475 | 5.5742E-01 | lmo1923 (aroE) | VmLPLQQmG | 2.556E-13 |
| 15023 | .....**VMASDYTRG**LSIRAIELVFEnL..... | 0.25073 | 2.4351E-01 | lmo1634 | VMASDYTRG | 6.973E-12 |
| 15024 | .....YMEKHS**VSRLVGAPP**..... | -0.4422 | 1.8422E-02 | lmo2206 (clpB) | VSRLVGAPP | 3.965E-12 |
| 15025 | .....KTASG**IVLPDSAKE**KPQSGKIVAVGSGRVLD.... | -0.49627 | 6.8168E-03 | lmo2069 (groES) | IVLPDSAKE | 4.253E-11 |
| 15026 | .....E**SVYLSAQKE**nPRLFAKIETF..... | 0.59506 | 8.1895E-04 | lmo1574 (dnaE) | SVYLSAQKE | 2.233E-11 |
| 15027 | .....RDNYL**FQLYEATGT**PIHPmSPF..... | -0.05049 | 8.7162E-01 | lmo2712 | FQLYEATGT | 5.383E-11 |

[a]Peptide modifications: oxidized Met (m), deamidated-N (n), and acetyl (k or n-term).

[b]Log2 Fold Change Listeria 6 hr versus 10 min.
[c]Cutoff:alpha=0.00005
[d]Score based on conformation to I-Ab-binding motif. Max=$9.9E^{-8}$. Cutoff=$9.0E^{-11}$.

**Table 2:** *Listeria* epitopes predicted by BOTA

| SEQ ID NO: | BOTA Prediction (*Listeria*) | gene name | BOTA score | Motif score | Peptidomics Hit |
|---|---|---|---|---|---|
| 15028 | ....ERWNEKYAQAYPNVSAKID ... | Imo0202 (hly) (LLO) | 0.557556 | 9.728E-11 | .....WNEKYAQAYPNVS..... |
| 15029 | ....ETQHYYGLPVADSAIDRGPL ... | lmo2558 (ami) | 0.675373 | 6.385E-11 | ...APGQETQHYYGLPVADSAIDR. ..... |
| 15030 | ....FRYVFDTAKATAASSYPGS ... | lmo2185 | 0.760419 | 1.932E-10 | .....ADFRYVFD TAKATAAS SYPG..... |
| 15031 | ....DTTVKFTLPTVAPAFENT IKT.. | lmo0135 | 0.532495 | 5.607E-09 | ...AVDDTTVKFTLPTVAPAFENT ..... |
| 15032 | .....EKWATPVSNVSTSSATSS .. | lmo0186 | 0.616773 | 4.923E-10 | .....GTKEKWATPVSNVSTSSA ..... |
| 15033 | ....INQAYTGS TALSD GLNKM.... | lmo2360 | 0.481871 | 7.495E-11 | .....GGINQAYTGSTALSDGLN ..... |
| 15034 | ...GIAVTAFAAPT IASAS TVW .... | lmo0582 (iap) | 0.671301 | 8.274E-09 | .....IAVTAFAAPTIASA. ..... |
| 15035 | ... AKE LVDNAPKALKE GIA..... | lmo0251 (rpll) | 0.744982 | 1.319E-10 | .....EAKE LVDNAPKALKE GIA ..... |
| 15036 | ...TVPAKSKTYPVYINEFALED .... | lmol927 (aroB) | 0.650328 | 1.187E-10 | .....TVRAKSKTYPVYInEFALE ..... |
| 15037 | ....YHGPIFLNASLSGVLE. ..... | lmo0201 (plcA) | 0.544155 | 6.548E-11 | N/A |
| 15038 | ....FGTHNVSSLSSGALNVHA .... | lmo0551 | 0.367349 | 5.869E-11 | N/A |
| 15039 | ....ANGDLYNAETSQYLGRL ..... | lmo0682 | 0.389302 | 5.763E-11 | N/A |
| 15040 | .....YELPKLPYTYDALE ..... | sod | 0.463634 | 8.126E-11 | N/A |
| 15041 | YIKKYDGDLKLALAAYNAG ..... | lmo0717 | 0.537408 | 1.079E-10 | N/A |
| 15042 | .FLITMAPEFPYLKPGSA. ..... | lmol883 | 0.547697 | 9.833E-11 | N/A |
| 15043 | .....PKSTVSSATSTAQ ..... | lmo0551 | 0.445178 | 1.292E-10 | N/A |
| 15044 | .....YPYVFGGSSPSTS ..... | lmo1104 | 0.517277 | 5.014E-10 | N/A |

**Materials and Methods**

**[0220] Immuno-affinity purification of MHCII complexes.** Bone marrow was harvested from C57/BL6 (WT), *Atgl6ll* $^{f/f}$ x CD1 lc-Cre (Conway et al., 2013b), and *Atg5*$^{f/f}$ x CDl lc-Cre (Conway et al., 2013a) mice. BMDCs were differentiated from bone marrow for seven days in antibiotic-free DMEM supplemented with 20% FBS and GM-CSF (2% conditioned TOPO medium). In parallel, *L. monocytogenes* strain EGDe (ATCC BA-679) was cultured overnight in BHI medium, washed in PBS, and cocultured with BMDCs at an MOI of approximately 100: 1 for 30 minutes. Tissue culture dishes containing cocultures were washed in PBS and cultured for an additional 10 minutes or 6 hours in DC media containing gentamicin (30 μg/ml). BMDCs were then harvested by scraping, washed in PBS, and lysed in 1% NP-40, 4 mM $MgCl_2$, 6 μg/ml DNaseI from bovine pancreas (Sigma Aldrich), and PBS pH 7.4 at 250 x $10^6$ cells/4ml lysis buffer/sample. Clarified lysates were subjected to immunoprecipitation overnight at 4°C with gentle rotation. Immunoprecipitation was performed with NHS Mag Sepharose (GE Healthcare) covalently coupled to anti-H2-IA (clone Y3P (Janeway et al., 1984)). Each sample contained 130 μl packed beads and approximately 650 μg antibody. Beads were then washed twice in PBS containing 0.1% NP-40 and three times in PBS.

**[0221] MHC-peptide elution and desalting.** Peptides were eluted from MHC complexes and desalted on in-house-built Empore C18 StageTips (3M, 2315) as described previously (Rappsilber et al., 2007). Sample loading, washes, and elution were performed on a tabletop centrifuge at a maximum speed of 2,000-3,500 x *g*. Briefly, StageTips were equilibrated with 2 × 100 μl washes of methanol, 2 × 50 μl washes of 50% acetonitrile/0. 1% formic acid (FA), and 2 × 100 μl washes of 1%, FA. In a tube, the dried beads from MHC-associated peptide IPs were thawed at 4°C, reconstituted in 50 μl 3%ACN/5% FA, and loaded onto StageTips. The beads were washed with 50 μL 1%, FA, and peptides were further eluted using two rounds of 5-minute incubations in 10% acetic acid. The combined wash and elution volumes were combined and loaded onto StageTips. The tubes containing the IP beads were washed again with 50 μl 1% FA, and this volume was also loaded onto StageTips. Peptides were washed twice on the StageTip with 100 μl 1% FA. Peptides were eluted using a step gradient of 20 μL 20%ACN/0.1% FA, 20 μL 40%ACN/0.1% FA, and 20 μl 60%ACN/0.1% FA. Step elutions were combined and dried to completion.

**[0222] iTRAQ 4-plex labeling for quantitative proteomics.** Quantitative proteomics was performed as described previously (Lassen et al., 2014; Mertins et al., 2014). Briefly, each peptide

mixture was reconstituted in 20 μL dissolution buffer labeled with 0.5 units (40 μL of 80 uL iTRAQ4 reagent in ethanol) of iTRAQ 4 reagent for 1 hour at room temperature (RT) (~22°C). Excess reagent was quenched with 5 μï⸗ Tris/HCl for 30 minutes at RT. iTRAQ4-plexes were combined, dried to completion, and acidified by adding 1% FA (150 μL). Samples were desalted on in-house built Empore SCX/C18 StageTips (3M, 2315, 2251) as described previously (Rappsilber et al., 2007). Peptides were eluted from the C18/SCX StageTips with two pH cuts (5.5, 11), and the remaining 20% ACN was diluted with 1% FA. Samples were then loaded on in-house built Empore CI8 StageTips, desalted, and dried to completion as described above.

**[0223] MHC-peptide sequencing by tandem mass spectrometry.** All nanoLC-ESI-MS/MS analyses employed the same LC separation conditions described below. Samples were chromatographically separated using a Proxeon Easy NanoLC 1000 (Thermo Scientific, San Jose, CA) fitted with a PicoFrit (New Objective, Woburn, MA) 75-μπι inner diameter capillary with a 10-μπι emitter was packed under pressure to -20 cm with of C18 Reprosil beads (1.9 μm particle size, 200 **A** pore size, Dr. Maisch GmBH) and heated at 50°C during separation. Samples were reconstituted in 9 μL 3%ACN/5% FA, and 3 μL (- IOOχIO$^6$ cell equivalents) was injected for analysis. Peptides were eluted with a linear gradient from 7-30% of Buffer B (0. 1% FA/90% ACN) over 82 min, 30-90% Buffer B over 6 min, and then held at 90% Buffer B for 15 min at 200 nL/min (Buffer A: 0.1% FA/3% ACN) to yield ~11 sec peak widths. During data-dependent acquisition, eluted peptides were introduced into a Q-Exactive FIF mass spectrometer (Thermo Scientific) equipped with a nanoelectrospray source (James A. Hill Instrument Services, Arlington, MA) at 2.15 kV. A full-scan MS was acquired at a resolution of 60,000 from 300 to 1,800 *m/z* (AGC target le6, 20ms Max IT). Each full scan was followed by top 15 data-dependent MS2 scans at resolution 15,000, using an isolation width of 1.7 m/z with a 0.3 m/z offset, a fixed first mass at 100 *m/z,* a collision energy of 29, an ACG Target of 5e4, and a max fill time of 100 ms max injection time. An isolation offset of 0.3 m/z was used so that doubly charged precursor isotope distributions would be centered in the isolation window. Some MHC-associated peptides tend to be short (<15 amino acids) so the monoisotopic peak is nearly always the tallest peak in the isotope cluster and the mass spectrometer acquisition software places the tallest isotopic peak in the center of the isolation window in the absence of a specified offset. Dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 10 sec. Charge state screening was enabled along

with monoisotopic precursor selection using Peptide Match Preferred to prevent triggering of MS/MS on precursor ions with charge state 1, >6, or unassigned.

**[0224] Interpretation of MS/MS data.** Mass spectra were interpreted using the Spectrum Mill software package v5.1 pre-Release (Agilent Technologies). MS/MS spectra were excluded from searching if they did not have a precursor MH+ in the range of 600-4000, had a precursor charge > 5, or had a minimum of < 5 detected peaks. Merging of similar spectra with the same precursor m/z acquired in the same chromatographic peak was disabled. High resolution MS/MS spectra were searched against a UniProt database containing reference proteome sequences (including isoforms and excluding fragments) from human mouse (41,157 entries) with a set of common laboratory contaminant proteins (150 sequences) to yield a total of 41,307 redundant sequences. The sequences were downloaded from the UniProt website in April 2013.

**[0225]** Prior to both search rounds all MS/MS had to pass the spectral quality filter with a sequence tag length > 3, i.e., minimum of 4 masses separated by the in-chain mass of an amino acid. In the first round search, all spectra were searched using a no-enzyme specificity, fixed modification of cysteine as unmodified, fixed modification of partial iTRAQ labeling, the following variable modifications (oxidized methionine, deamidation, acetyl-N-term), a precursor mass tolerance of ±10 ppm, product mass tolerance of ±20 ppm, and a minimum matched peak intensity of 50%. Peptide spectrum matches (PSMs) for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to apply target-decoy based FDR estimation at the PSM level to set scoring threshold criteria. Peptide autovalidation was performed with an auto thresholds strategy using a minimum sequence length of 7, automatic variable range precursor mass filtering, and score and delta Rankl - Rank2 score thresholds optimized across all LC-MS/MS runs. This yielded a PSM level FDR estimate for precursor charges 1 thru 7 of < 1.0% for each precursor charge state. In the second round search, all remaining spectra that that were not confidently identified in the first round were searched using the above parameters against a RefSeq database containing *L. monocytogenes EGD-e* reference protein sequences (2,867) downloaded in December 2014. An additional round of FDR thresholding as described above was applied to PSMs from the second round search to estimate FDR by species (mouse vs. *L. monocytogenes EGD-e).* The combined PSMs from each round had a peptide level FDR < 2.0%. Only *L. monocytogenes EGD-e* peptides that did not overlap with human peptides were reported.

**[0226]** ***Listeria*** **infection and ELISPOT.** *L. monocytogenes* strain EGDe (ATCC BA-679) was grown to log phase in BHI, washed in PBS, and used to inoculate mice by i.p. injection with a dose of $1 \times 10^4$ CFU/200 μï/mouse. Seven days after infection, spleens were harvested, red blood cells were lysed, and single cell suspensions were replated at $2.5 \times 10^5$/200 μï/well on 96-well ELISPOT plates (Millipore MAIPS4510) precoated with anti-IFNy (BD Pharmingen 551881) at a concentration of 5 μg/ml in PBS. Peptides were added at a concentration of 100 nM to restimulate splenocytes overnight. After stimulation, plates were washed, blocked, and incubated with biotinylated anti-IFNγ (BD Pharmingen 551881) at a concentration of 2 μg/ml and streptavidin-AKP (BD Pharmingen 554065) at a dilution of 1:1000. After 60 minutes, plates were washed and developed with 3-amino-9-etyhlcarbazole (AEC, Sigma-Aldrich, A6926). IFNy-secreting cells were scanned and enumerated using an Immunospot Analyzer ELISpot reader (CTL). All samples were run in triplicate. The following peptides were synthesized by New England Peptide: lmo2185_293-312 ADFRYVFDTAKATAASSYPG (SEQ ID NO: 14986), lmo0202_l 89-204 WNEKYAQAYPNVSAKI (SEQ ID NO: 14981), lmo0135_150-169 VDDTTVKFTLPTVAPAFENT (SEQ ID NO: 14997), lmo2558_533-553 APGQETQHY YGLPVAD SAIDR (SEQ ID NO: 14984), lmo2360_289-306 GGINQAYTGSTALSDGLN (SEQ ID NO: 15009), lmo0186_285-303 GTKEKVVATPVSNVST SSA (SEQ ID NO: 15005), lmo0582a_26-46 STVVVEAGDTLWGIAQSKGTT (SEQ ID NO: 15012), lmo0582b_12-25 IAVTAFAAPTIASA (SEQ ID NO: 15013). To calculate the frequency of $CD4^+$ cells in each splenocyte sample from infected mice, an aliquot of cells from each mouse was stained for flow cytometric analysis. $1 \times 10^6$ cells were incubated with 2.4G2 Mouse Fc block in PBS/FBS (BD Pharmingen, 553142) for 20 min at 4°C. Cells were then washed and stained with FITC-conjugated anti-mouse CD4 (Biolegend, 100510) for 20 min at 4°C. Fluorescently labeled cells were acquired on the FACSVerse flow cytometer (BD Biosciences) and analyzed using FlowJo Analysis Software (Tree Star).

**[0227]** **BOTA algorithm architecture.** BOTA starts with an input genome and the associated genome annotation. The genome annotation may be in GFF3 format. It first extracts the amino acid sequences of the protein-coding genes and then performs the following analysis: (i) domain identification using hmmscan of HMMER3.1b2 (Eddy, 201 1) against (Finn et al., 2016); (ii) cellular localization prediction using PSORTb v3.0.2 (Yu et al., 2010) with default settings; and

(iii) transmembrane topology prediction by HMMTOP (Tusnady and Simon, 1998). This information is then integrated to generate a list of candidate peptides following three criteria: (i) the protein should be located in the outer membrane, cell wall (if applicable), or extracellular space; (ii) if located in the outer membrane or cell wall, only the outfacing part of the protein will be considered; and (iii) the peptide should present sufficient accessibility as decided by three rules: (iii-a) it cannot be fewer than eight amino acids away from anchoring domain in the cell wall or outer membrane; (iii-b) it cannot be located in domains shorter than 30 amino acids due to the fact that small domains are usually tightly folded; and (iii-c) it cannot be flanked by two domains that are fewer than 20 amino acids apart. These candidate peptides are then scored by the deep neural network as described below. For each peptide classified as a candidate epitope, BOTA further validates it with the motif score as defined previously, and a randomized score. For the motif score validation, it calculates all the 9-mers within the peptide and requires the maximum to be larger than $5 \times 10^{-11}$; for the randomized score, BOTA shuffles the amino acids in the 9-mer and calculates the motif score for each of the randomized 9-mers. The original 9-mer' s motif score requires a rank in the top 30% of all randomized 9-mers.

**[0228]** In certain example embodiments, protein intracellular prediction may be achieved by using a lowest common ancestor (LCA) database to train a convolutional neural network (CNN) model, a generative adversarial network (GAN) models, or both. In other example embodiments, multiple transmembranal databases may be used to train a domain prediction model for use in conjunction with, or as an alternative to HMMTOP.

**[0229]** **Deep neural network for MHCII binder prediction.** Applicants employed a deep neural network scheme to develop an MHCII-binder predictor. In brief, Applicants first encoded every amino acid into a ^-dimensional binary vector $b$, with half of its elements being 1 and the rest being 0, chosen at random. Therefore, given a peptide with length / longer than $k$, it is first converted to an $l \times b$ descriptor matrix $S$, in which $S_{ij}$=1 if the ż-th amino acid's 1-valued indices overlap with $j$, otherwise Sy=0. The matrix $S$ is then normalized by row sum to become $S'$ such that

$$S'_{ij} = S_{ij} / \sum_{j=1}^{b} S_{ij}$$

**[0230]** $S'$ is then convoluted into an $(l\text{-}k+1) \times d$ matrix $X$, where $d$ is the number of pre-trained motif network models within the overall model and $k$ is the length of such binding cores. $X_j$

represents the score of motif network model $j$ aligned to position $i$. The cohort of motif network models are arranged in a $d \times k \times b$ array $H$ with $H_{ijk}$ being the *d-t*h motif network model aligned to the &-th position of the *b-t*h set of amino acids **(Fig. 8D).** This sequence conversion and convolution setup is similar to model developed by Alipanahi et al (Alipanahi et al., 2015).

**[0231]** With the convoluted matrix $X$, Applicants filter it with a max-rectified linear unit (ReLU) layer, and the rectified matrix $Y$ is then fed into a max pooling stage to be transformed into <i-dimensional vector Z, in which

$$zj = \max(Y_1j, Y_2j, \ldots, Ynj)$$

**[0232]** This <i-dimensional vector Z is then used as input for a neural network with Maxout dropout model. Z is then used as the input for a standard output layer for final prediction calculation.

**[0233]** The goal is to minimize the prediction error as measured by 1-norm distance of all the peptides. Back propagation with a stochastic gradient descent method using mini batch size at 64 was used to reach optimal weights. To train the weights, the mouse epitope peptides were used as training set true positive; Applicants also added *in silico* true negative peptides by surveying an equal number of the peptides that are not part of the peptidomics data readout at random. Applicants first randomly constructed 100 replicates of 3-fold cross validation datasets using the epitopes and the *in silico* true negative sequences. For each replicate, BOTA's initial state weights were assigned randomly and then trained until performance plateaued (<0.1% improvement in ten iterations). The validation's average AUC was used as a measure of model fitness; and the one with the highest average AUC among the 100 replicates was chosen to generate the final optimal parameters by using all mouse peptides **(Fig. 9C).** Then the optimal model constructed from the previous step were used to predict the unique *Listeria* peptides, and netMHCIIpan was used to predict the affinity of the same set of peptides with default settings.

**[0234]** **TCRseq and 5' digital gene expression (DGE).** 5'DGE RNAseq data were analyzed using an in-house data analysis pipeline. Reads were aligned to the mm10 genome using bwa (Li et al., 2009). The downstream analysis was carried out using the Seurat package (www.satijalab.org/seurat). The dataset was filtered to remove genes that were expressed in less than 10% of the data or cells that expressed transcripts that mapped to fewer than 500 unique genes. Further, to remove doublet cells, cells that displayed more heterogeneity than the 90$^{th}$ quantile (1846 unique genes) were filtered out. Highly variable, highly expressed genes (log(o^)

> 0.75 and log^) > 1.5) were identified from a mean-variance dispersion analysis. These genes were then used for clustering the single cells using the SNN-Cliq algorithm, which maps the cells onto a k-nearest neighbor graph and then finds "cliques" of cells expressing similar genes (McDavid et al., 2014; Xu and Su, 2015). It was found that cells expressing Ifng and Tbx21 belonged primarily to cluster 0. Cluster identities of Thl7 single cells were overlaid on a t-SNE plots. The cells were clustered on the basis of highly expressed, highly varying genes, using a graph-based method, SNN-Cliq (McDavid et al., 2014). Violin/Feature plot: Ifng is identified as a marker gene for cluster 0 based on the bimod test for differential expression (Xu et al., 2015).

**[0235]** **TCR Screening.** TCR-negative BW5 147 cells were obtained from ATCC. The stable sub-line BW_4-28 was produced by introduction of murine CD4 and CD28 by lentiviral transduction. The Orf encoding CD4-P2A-CD28 was synthesized (IDT) and cloned in place of spCas9-P2A-BlastR into pXPR_BRDlOl (Genetic Perturbation Platform, Broad Institute) by Gibson assembly. The stable sub-line BW_B7/4 was produced by introduction of chimeric murine CD86 cytoplasmic domain fused in-frame with CD4 transmembrane and cytoplasmic domains by lentiviral transduction. The Orf encoding CD86/CD4 was synthesized (IDT) and cloned in place of spCas9-P2A-BlastR into pXPR_BRDlOl (Genetic Perturbation Platform, Broad Institute) by Gibson assembly.

**[0236]** scTCRz_v3 vector was derived from pLX_TRC307 (Genetic Perturbation Platform, Broad Institute) by replacing the stuffer sequence with the following codon-optimized Orf: mTrbcl-hCD3zeta(transmembrane and cytoplasmic domains)-P2A-hIgKleader-mTrac hCD3zeta(transmembrane and cytoplasmic domains). Single-chain TCRs were synthesized as gBlocks (IDT) comprising Trav-Traj-linker(3x$G_4$SGGGG)-Trbv-Trbj and cloned by Gibson assembly into Nhel and BsrGI sites upstream of Trbcl in scTCRz_v3 vector. MHCIIz_v2 vector was derived from pLX_TRC307 (Genetic Perturbation Platform, Broad Institute) by replacing the stuffer sequence with the following sequence: H2Ab1leader-XmaI-linker(3x$G_4$S)-H2Abl (extracellular and transmembrane domains)-hCD3zeta(cytoplasmic domain)-P2A H2Aa(extracellular and transmembrane domains)-hCD3zeta(cytoplasmic domain). Peptide epitope-encoding sequences were synthesized as ultramers (IDT) and cloned by Gibson assembly into the Xmal site in the MHCIIz_v2 vector. pMHCII and scTCRs were introduced into BW_B7/4 and BW_4-28 cells, respectively, by lentiviral transduction and maintained under puromycin selection (3ug/ml). To define antigen specificity of TCRs, BW_4-28 cells expressing scTCRs were

cocultured with HEK 293T cells transfected with pMHCII and CD86/CD4 or BW_B7/4 cells expressing pMHCII. Cells were combined at a 1:1 ratio in 96-well flat-bottom plates (total 100,000 cells/well). Culture supernatants were harvested after 18 hours, and IL-2 was measured by cytokine bead array (BD Pharmingen) or cells were harvested after 18 hours, and mRNA expression of the activation marker Nur77 (Nr4al) was quantified relative to Actb.

**[0237]** **Animals.** Mice were bred in specific-pathogen-free facilities at Massachusetts General Hospital (Boston, MA) prior to infection and transferred to biosafety level 2 (BL2) housing upon *L. monocytogenes* infection. All animal studies were conducted in compliance with ethical regulations and were approved by the Institutional Care and Use Committee (IACUC) at Massachusetts General Hospital. *Atg5*$^{f/f}$ and *Atgl6ll* $^{f/f}$ mice were generated as previously described (Conway et al., 2013b; Hara et al., 2006) and bred with mice expressing *Cre* recombinase under the control of the CDl lc promoter (CDl lc-Cre) (Jackson Laboratories). All mice were maintained on food and water *ad libitum,* used between 8-12 weeks of age, and age- and sex-matched for each experiment.

**[0238]** **Statistical analysis.** For data analysis, iTRAQ 4 ratios for the two biological replicates were filtered to retain only those deemed reproducible as described previously (Kronke et al., 2015). Reproducibility was based on replicates being confined within the 95% limits of agreement of a Bland-Altman plot (Bland and Altman, 1986). Reproducible replicates were then subjected to a moderated *t*-testto assess statistical significance (Smyth, 2004). To generate the I-A$^b$-binding motif, all mouse peptides were aligned by MultAlin (Corpet, 1988) and visualized with WebLogo (Crooks et al., 2004). The frequencies of each amino acid at each position of the 9-mer core were used to generate a position-weight matrix. I-A$^b$ binding scores for *Listeria* peptides were produced from this matrix and represented as the product of amino acid frequencies for each position of the 9-mer core (Table 1). Epitope mapping and protein domain overlay were performed with NCBFs Conserved Domain Database (Marchler-Bauer et al., 2015). Subcellular localization of antigenic *Listeria* proteins was predicted by PSORTb v3.0.2 (Yu et al., 2010), and by COMPARTMENTS (Binder et al., 2014) for mouse proteins. Transmembrane topology was predicted with HMMTOP (Tusnady and Simon, 1998).

**REFERENCES**

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol *33,* 831-838.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31,* 166-169.

Andreatta, M., Schafer-Nielsen, C , Lund, O., Buus, S., and Nielsen, M. (201 1). NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. PLoS ONE *6,* e26781.

Arunachalam, B., Phan, U.T., Geuze, H.J., and Cresswell, P. (2000). Enzymatic reduction of disulfide bonds in lysosomes: characterization of a gamma-interferon-inducible lysosomal thiol reductase (GILT). Proc Natl Acad Sci U S A *97,* 745-750.

Babbitt, B.P., Allen, P.M., Matsueda, G., Haber, E., and Unanue, E.R. (1985). Binding of immunogenic peptides to la histocompatibility molecules. Nature *317,* 359-361.

Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C , O'Donoghue, S.I., Schneider, R., and Jensen, L.J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford) *2014,* bau012.

Bland, J.M., and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet *1,* 307-310.

Chatterjee, S.S., Hossain, H., Otten, S., Kuenne, C , Kuchmina, K., Machata, S., Domann, E., Chakraborty, T., and Hain, T. (2006). Intracellular gene expression profile of Listeria monocytogenes. Infect Immun *74,* 1323-1338.

Chicz, R.M., Urban, R.G., Gorga, J.C., Vignali, D.A., Lane, W.S., and Strominger, J.L. (1993). Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. J Exp Med 775, 27-47.

Chicz, R.M., Urban, R.G., Lane, W.S., Gorga, J.C., Stern, L.J., Vignali, D.A., and Strominger, J.L. (1992). Predominant naturally processed peptides bound to HLA-DRl are derived from MHC-related molecules and are heterogeneous in size. Nature *358,* 764-768.

Christmann, B.S., et al. Human seroreactivity to gut microbiota antigens. J. Allergy Clin. Immunol. 136, 1378-1386 el371-1375 (2015).

Conway, K.L., Kuballa, P., Song, J.H., Patel, K.K., Castoreno, A.B., Yilmaz, O.H., Jijon, H.B., Zhang, M., Aldrich, L.N., Villablanca, E.J., *etal.* (2013a). Atgl611 is required for autophagy

in intestinal epithelial cells and protection of mice from Salmonella infection. Gastroenterology *145,* 1347-1357.

Conway, K.L., Kuballa, P., Khor, B., Zhang, M., Shi, H.N., Virgin, H.W., and Xavier, R.J. (2013b). ATG5 regulates plasma cell differentiation. Autophagy *9,* 528-537.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res *16,* 10881-10890.

Cresswell, P., Arunachalam, B., Bangia, N., Dick, T., Diedrich, G., Hughes, E., and Marie, M. (1999). Thiol oxidation and reduction in MHC-restricted antigen processing and presentation. Immunol Res *19,* 191-200.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14,* 1188-1 190.

Depontieu, F.R., Qian, J., Zarling, A.L., McMiller, T.L., Salay, T.M., Norris, A., English, A.M., Shabanowitz, J., Engelhard, V.H., Hunt, D.F., *et al.* (2009). Identification of tumor-associated, MHC class Il-restricted phosphopeptides as targets for immunotherapy. Proc Natl Acad Sci U S *A 106,* 12073-12078.

Dongre, A.R., et al. In vivo MHC class II presentation of cytosolic proteins revealed by rapid automated tandem mass spectrometry and functional analyses. Eur. J. Immunol. 31, 1485-1494 (2001).

Eddy, S.R. (201 1). Accelerated Profile HMM Searches. PLoS Comput Biol 7, el002195.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res *44,* D279-285.

Hall, A.B., Tolonen, A.C. & Xavier, R.J. Human genetic variation and the gut microbiome in disease. Nat. Rev. Genet. 18, 690-699 (2017).

Hara, T., Nakamura, K., Matsui, M., Yamamoto, A., Nakahara, Y., Suzuki-Migishima, R., Yokoyama, M., Mishima, K., Saito, L, Okano, H., *et al.* (2006). Suppression of basal autophagy in neural cells causes neurodegenerative disease in mice. Nature *441,* 885-889.

Heng, T.S., Painter, M.W. & Immunological Genome Project, C. The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. 9, 1091-1094 (2008).

Hsieh, C.S., deRoos, P., Honey, K., Beers, C , and Rudensky, A.Y. (2002). A role for cathepsin L and cathepsin S in peptide generation for MHC class II presentation. J Immunol *168,* 2618-2625.

Hsing, L.C., and Rudensky, A.Y. (2005). The lysosomal cysteine proteases in MHC class II antigen presentation. Immunol Rev *207,* 229-241.

Hunt, D.F., Michel, H., Dickinson, T.A., Shabanowitz, J., Cox, A.L., Sakaguchi, K., Appella, E., Grey, H.M., and Sette, A. (1992). Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. Science *256,* 1817-1820.

Janeway, C.A., Jr., Conrad, P.J., Lerner, E.A., Babich, J., Wettstein, P., and Murphy, D.B. (1984). Monoclonal antibodies specific for la glycoproteins raised by immunization with activated T cells: possible role of T cellbound la antigens as targets of immunoregulatory T cells. J Immunol *132,* 662-667.

Kim, A., and Sadegh-Nasseri, S. (2015). Determinants of immunodominance for CD4 T cells. Curr Opin Immunol *34,* 9-15.

Kim, D., Pertea, G , Trapnell, C , Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14,* R36.

Kronke, J., Fink, E.C., Hollenbach, P.W., MacBeth, K.J., Hurst, S.N., Udeshi, N.D., Chamberlain, P.P., Mani, D.R., Man, H.W., Gandhi, A.K., *et al.* (2015). Lenalidomide induces ubiquitination and degradation of CKlalpha in del(5q) MDS. Nature *523,* 183-188.

Lassen, K.G., Kuballa, P., Conway, K.L., Patel, K.K., Becker, C.E., Peloquin, J.M., Villablanca, E.J., Norman, J.M., Liu, T.C., Heath, R.J., *et al.* (2014). Atgl6Ll T300A variant decreases selective autophagy resulting in altered cytokine signaling and decreased antibacterial defense. Proc Natl Acad Sci U S A *111,* 7741-7746.

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760 (2009).

Lippolis, J.D., White, F.M., Marto, J.A., Luckey, C.J., Bullock, T.N., Shabanowitz, J., Hunt, D.F., and Engelhard, V.H. (2002). Analysis of MHC class II antigen processing by quantitation of peptides that constitute nested sets. J Immunol *169,* 5089-5097.

Liu, X., et al. Alternate interactions define the binding of peptides to the MHC molecule IA(b). Proc. Natl. Acad. Sci. U. S. A. 99, 8820-8825 (2002).

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., *et al.* (2015). CDD: NCBI's conserved domain database. Nucleic Acids Res *43,* D222-226.

McDavid, A., et al. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. PLoS Comput. Biol. 10, el003696 (2014).

Mertins, P., Yang, F., Liu, T., Mani, D.R., Petyuk, V.A., Gillette, M.A., Clauser, K.R., Qiao, J.W., Gritsenko, M.A., Moore, R.J., *et al.* (2014). Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. Mol Cell Proteomics *13,* 1690-1704.

Miyazaki, T., Wolf, P., Tourne, S., Waltzinger, C , Dierich, A., Barois, N., Ploegh, H., Benoist, C , and Mathis, D. (1996). Mice lacking H2-M complexes, enigmatic elements of the MHC class II peptide-loading pathway. Cell *84,* 531-541.

Mommen, G.P., Marino, F., Meiring, H.D., Poelen, M.C., van Gaans-van den Brink, J.A., Mohammed, S., Heck, A.J., and van Els, C.A. (2016). Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. Mol Cell Proteomics *15,* 1412-1423.

Nelson, C.A., Roof, R.W., McCourt, D.W., and Unanue, E.R. (1992). Identification of the naturally processed form of hen egg white lysozyme bound to the murine major histocompatibility complex class II molecule I-Ak. Proc Natl Acad Sci U S A *89,* 7380-7383.

Ormerod, K.L., et al. Genomic characterization of the 1 uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome 4, 36 (2016).

Palm, N.W., de Zoete, M.R. & Flavell, R.A. Immune-microbiota interactions in health and disease. Clin. Immunol. 159, 122-127 (2015).

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc *2,* 1896-1906.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26,* 139-140.

Rudensky, A., Preston-Hurlburt, P., Hong, S.C., Barlow, A., and Janeway, C.A., Jr. (1991). Sequence analysis of peptides bound to MHC class II molecules. Nature *353,* 622-627.

Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R.V., Widdowson, M.A., Roy, S.L., Jones, J.L., and Griffin, P.M. (2011). Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17, 7-15.

Schindelin, J., Arganda-Carreras, L, Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C , Saalfeld, S., Schmid, B., *etal.* (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods *9,* 676-682.

Schulze, M.S., and Wucherpfennig, K.W. (2012). The mechanism of HLA-DM induced peptide exchange in the MHC class II antigen presentation pathway. Curr Opin Immunol *24,* 105-111.

Seamons, A., Sutton, J., Bai, D., Baird, E., Bonn, N., Kafsack, B.F., Shabanowitz, J., Hunt, D.F., Beeson, C , and Goverman, J. (2003). Competition between two MHC binding registers in a single peptide processed from myelin basic protein influences tolerance and susceptibility to autoimmunity. J Exp Med *197,* 1391-1397.

Sette, A., Ceman, S., Kubo, R.T., Sakaguchi, K., Appella, E., Hunt, D.F., Davis, T.A., Michel, H., Shabanowitz, J., Rudersdorf, R., *et al.* (1992). Invariant chain peptides in most HLA-DR molecules of an antigen-processing mutant. Science *258,* 1801-1804.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3,* Article3.

Sofiron, A., Ritz, D., Neri, D., and Fugmann, T. (2016). High-resolution analysis of the murine MHC class II immunopeptidome. Eur J Immunol *46,* 319-328.

Stern, L.J., Brown, J.H., Jardetzky, T.S., Gorga, J.C., Urban, R.G., Strominger, J.L., and Wiley, D.C. (1994). Crystal structure of the human class II MHC protein HLA-DRl complexed with an influenza virus peptide. Nature *368,* 215-221.

Stoll, M.L., et al. Altered microbiota associated with abnormal humoral immune responses to commensal organisms in enthesitis-related arthritis. Arthritis Res. Ther. 16, 486 (2014).

Suri, A., Walters, J.J., Rohrs, H.W., Gross, M.L., and Unanue, E.R. (2008). First signature of islet beta-cell-derived naturally processed peptides selected by diabetogenic class II MHC molecules. J Immunol *180,* 3849-3856.

Tusnady, G.E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol *283,* 489-506.

Wang, P., Sidney, J., Dow, C , Mothe, B., Sette, A., and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol *4,* el000048.

Weber, K.S., et al. Distinct CD4+ helper T cells involved in primary and secondary responses to infection. Proc. Natl. Acad. Sci. U. S. A. 109, 951 1-9516 (2012).

Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 31, 1974-1980 (2015).

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., *et al.* (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics *26,* 1608-1615.

Zeng, M.Y., et al. GutMicrobiota-Induced Immunoglobulin G Controls Systemic Infection by Symbiotic Bacteria and Pathogens. Immunity 44, 647-658 (2016).

Zhu, Y., Rudensky, A.Y., Corper, A.L., Teyton, L. & Wilson, LA. Crystal structure of MHC class II I-Ab in complex with a human CLIP peptide: prediction of an I-Ab peptide-binding motif. J. Mol. Biol. 326, 1157-1 174 (2003).

**[0239]** The invention is further described by the following numbered paragraphs:

1. A method for identifying MHCII antigenic epitopes comprising:

generating, by a processor, a set of candidate antigens from one or more input genome sequences; and

generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network, wherein the deep neural network is trained using a set of MHCII-presented peptides isolated from antigen presenting cells.

2. The method of paragraph 1, wherein the antigen presenting cells are dendritic cells.

3. The method of paragraph 1, wherein generating a set of candidate antigens from one or more input genome sequences comprises:

predicting, by the processor, genes from an input genome sequence; and

defining, by the processor, a set of candidate antigens from the set of predicted genes based on one or more of protein cellular location, transmembrane structure, and domain distribution.

4. The method of paragraph 3, wherein defining, by the processor, a set of candidate antigens from the set of predicted genes, comprises:

a) selecting surface and secreted proteins;

b) masking intracellular regions and transmembrane domains of the surface and extracellular proteins;

c) excluding regions that are within domains of less than 30 amino acids;

d) excluding regions between a series of inaccessible domains; and

e) excluding inter-domain regions between a series of adjacent domains wherein the inter-domain regions are less than or equal to 20 amino acids.

5. The method of paragraph 4, further comprising masking up to 20 amino acids flanking the intracellular regions and transmembrane domains.

6. The method of paragraph 4, wherein 8 amino acids flanking the intracellular regions and transmembrane domains are masked.

7. The method of paragraph 4, wherein the surface protein is a cell wall protein.

8. The method of paragraph 4, wherein the surface protein is an outer membrane protein

9. The method of any of paragraphs 1 to 8, wherein the antigenic epitopes are derived from candidate antigens more than 20 amino acids away from a transmembrane domain.

10. The method of any of paragraphs 1 to 9, wherein the candidate antigens comprise a tertiary structure required for proteolytic enzyme accessibility.

11. The method of any of paragraphs 1 to 10, wherein the candidate antigens comprise a defined MHCII-binding motif in the primary amino acid sequence of the predicted genes.

12. The method of paragraph 11, wherein the defined MHCII-binding motif is the I-$A^b$-binding motif.

13. The method of any of paragraphs 1 to 12, wherein the antigenic epitopes comprise a tertiary structure required for accessibility to a MHCII binding groove.

14. The method of any of paragraphs 1 to 13, wherein generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network comprises:

a) encoding each amino acid of an input candidate antigen into a p-dimensional binary vector;

b) converting the binary vector to a descriptor matrix S;

c) convoluting the descriptor matrix S to a scoring matrix X; and

d) transforming the scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for a neural network.

15. The method of any of paragraphs 1 to 14, wherein the method further comprises:

isolating MHCII-peptide complexes from antigen presenting cells;

isolating the peptides from the MHCII-peptide complexes;

sequencing the isolated peptides; and

training the deep neural network using the set of isolated peptides.

16. The method of paragraph 15, wherein the antigen presenting cells are exposed to a target cell type or target pathogen.

17. The method of paragraph 15 or 16, wherein the antigen presenting cells are dendritic cells.

18. The method of any of paragraphs 1 to 17, wherein the input genome sequence is derived from a target cell type or target pathogen.

19. The method of paragraph 18, wherein the candidate antigens are expressed in the target cell type or target pathogen.

20. The method of any of paragraph 16 to 19, wherein the target cell type or target pathogen is selected from the group consisting of a bacterium, a virus, a protozoa, an allergen and a diseased cell.

21. The method of paragraph 20, wherein the diseased cell is a cancer cell.

22. The method of paragraph 21, wherein the cancer cell is obtained from a subject.

23. The method of paragraph 22, further comprising identifying non-silent tumor specific somatic mutations from the subject specific cancer cell, wherein the ranked set of antigenic epitopes are generated from said mutations.

24. The method of paragraph 20, wherein the diseased cell is associated with an autoimmune disease.

25. The method of paragraph 20, wherein the target pathogen comprises a bacteria belonging to a family selected from the group consisting of *Bacillus, Bartonella, Bordetella, Borrelia, Brucella, Campylobacter, Chlamydia and Chlamydophila, Clostridium, Corynebacterium, Enterococcus, Escherichia, Francisella, Haemophilus, Helicobacter, Legionella, Leptospira, Listeria, Mycobacterium, Mycoplasma, Neisseria, Pseudomonas, Rickettsia, Salmonella, Shigella, Staphylococcus, Streptococcus, Treponema, Ureaplasma, Vibrio,* and *Yersinia.*

26. The method of any of paragraphs 1 to 25, wherein the antigen presenting cell comprises a subject specific MHCII allele.

27. The method of any of paragraphs 1 to 26, wherein the antigen presenting cell comprises a human MHCII allele.

28. The method of any of paragraphs 1 to 27, wherein the one or more input genome sequences is obtained by whole genome or whole exome sequencing.

29. The method of any of paragraphs 1 to 20, further comprising detecting whether one or more the antigenic epitopes is present in a sample from a subject suffering from an infection, autoimmune disease, allergy, or cancer.

30. The method of any one of paragraphs 1 to 29, wherein the antigenic epitopes are about 9 to 20 amino acids in length.

31. The method of any of paragraphs 1 to 30, wherein the deep neural network is trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 1-14979.

32. The method of any of paragraphs 1 to 31, wherein the deep neural network is trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 14980-15027.

33. The method of any of paragraphs 1 to 32, further comprising preparing a vaccine composition comprising one or more antigenic epitopes from the set of antigenic epitopes generated by the processor.

34. The method of paragraph 33, wherein the vaccine composition is directed to a bacterium, a virus or a protozoa.

35. The method of paragraph 33, wherein the vaccine composition is directed to a cancer.

36. The method of paragraph 33, wherein the vaccine composition is directed to an autoimmune disease, whereby administration of the vaccine induces tolerance to the antigenic epitope.

37. The method of paragraph 34, wherein the vaccine composition is directed to *Listeria.*

38. The method of paragraph 37, wherein the vaccine composition comprises one or more *Listeria* peptides selected from the peptides listed in Table 1.

39. The method of paragraph 38, wherein the one or more *Listeria* peptides are derived from lmo0202, lmo2558, lmo2185, or lmo0135.

40. The method of paragraph 39, wherein the one or more *Listeria* peptides are selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984 (APGQETQHYYGLPVADSAIDR), SEQ ID NO: 14986 (ADFRYVFDTAKATAASSYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT).

41. The method of any of paragraphs 33 to 40, wherein the vaccine comprises autologous dendritic cells or antigen presenting cells pulsed with the one or more peptides.

42. A method of identifying immunodominant epitopes comprising:

a) expressing in a first population of immune cells a peptide MHCII library comprising antigenic epitopes identified according to the method of any of paragraphs 1 to 32, wherein the first population of immune cells comprise a detectable reporter gene;

b) expressing in a second population of CD4+ immune cells a TCRab library expressing TCRab pairs identified in a subject suffering from an infectious disease, autoimmunity, cancer or allergy;

c) incubating the first population of immune cells with the second population of immune cells;

d) sorting immune cells positive for the detectable reporter gene; and

e) identifying peptides bound to MHCII in immune cells positive for the detectable reporter gene.

43. The method of paragraph 42, wherein the TCRab pairs are identified by single cell profiling.

44. The method of paragraphs 43, wherein TCRab pairs are determined by targeted single cell RNA-seq (TCRseq).

45. The method of paragraphs 43 or 44, wherein T cells are analyzed by RNA-seq to determine single cells that are activated and TCRab pairs are identified in the activated T cells.

46. The method of any of paragraphs 42 to 45, wherein the reporter is an inducible fluorescent marker protein, wherein the marker protein is induced when a TCRab pair detects a MHCII epitope.

47. The method of paragraph 46, wherein the fluorescent marker is under control of the IL-2 promoter.

48. The method of any of paragraphs 42 to 47, further comprising administering to the subject a vaccine comprising one or more of the immunodominant epitopes.

49. The method of paragraph 48, wherein the vaccine is a protective vaccine or a tolerizing vaccine.

50. The method of paragraphs 48 or 49, wherein the vaccine comprises autologous dendritic cells or antigen presenting cells pulsed with one or more of the immunodominant epitopes.

51. A method of identifying peptide antigens from a live bacterial pathogen comprising:

isolating MHCII-peptide complexes from dendritic cells exposed to a bacterial pathogen;

isolating the peptides from the MHCII-peptide complexes; and

sequencing the isolated peptides.

52. A peptide set for training a MHCII neural network comprising one or more of SEQ ID NOs: 1-14979.

53. A peptide set for training a MHCII neural network comprising one or more of SEQ ID Nos: 14980-15027.

54. A *Listeria* vaccine comprising one or more peptides selected from the peptides listed in Table 1.

55. The *Listeria* vaccine of paragraph 42, wherein the one or more peptides is derived from lmo0202, lmo2558, lmo2185, or lmo0135.

56. The *Listeria* vaccine of paragraph 43, wherein the one or more *Listeria* peptides is selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID

NO: 14984 (APGQETQHYYGLP VADSAIDR), SEQ ID NO: 14986 (ADFRYVFDTAKATAASSYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT).

57. The *Listeria* vaccine of any of paragraphs 42 to 44, wherein the vaccine comprises autologous dendritic cells or antigen presenting cells pulsed with the one or more peptides.

* * *

**[0240]** Various modifications and variations of the described methods, pharmaceutical compositions, and kits of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific embodiments, it will be understood that it is capable of further modifications and that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the art are intended to be within the scope of the invention. This application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure come within known customary practice within the art to which the invention pertains and may be applied to the essential features herein before set forth.

## CLAIMS

What is claimed is:

1. A method of preparing one or more peptides for an immunological composition comprising:

a) generating, using one or more processors, a set of candidate antigens from one or more input genome sequences;

b) generating, using the one or more processors, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network; and

c) formulating an immunological composition comprising one or more of the identified epitopes.

2. The method of claim 1, wherein the one or more input genome sequences is derived from a target pathogen, a commensal microorganism, or a diseased cell.

3. The method of claim 1, wherein the deep neural network is trained using a set of MHCII-presented peptides bound to antigen presenting cells.

4. The method of claim 3, wherein the antigen presenting cells are dendritic cells.

5. The method of claim 1, wherein generating a set of candidate antigens from one or more input genome sequences comprises:

predicting, by the processor, genes from an input genome sequence; and

defining, by the processor, a set of candidate antigens from the set of predicted genes based on one or more of protein cellular location, transmembrane structure, and domain distribution.

6. The method of claim 5, wherein defining, by the processor, a set of candidate antigens from the set of predicted genes, comprises:

a) selecting surface and secreted proteins;

b) masking intracellular regions and transmembrane domains of the surface and extracellular proteins;

c) excluding regions that are within domains of less than 30 amino acids;

d) excluding regions between a series of inaccessible domains; and

e) excluding inter-domain regions between a series of adjacent domains wherein the inter-domain regions are less than or equal to 20 amino acids.

7. The method of claim 6, further comprising masking up to 20 amino acids flanking the intracellular regions and transmembrane domains.

8. The method of claim 6, wherein 8 amino acids flanking the intracellular regions and transmembrane domains are masked.

9. The method of claim 6, wherein the surface protein is a cell wall protein.

10. The method of claim 6, wherein the surface protein is an outer membrane protein

11. The method of claim 1, wherein the antigenic epitopes are derived from candidate antigens more than 20 amino acids away from a transmembrane domain.

12. The method of claim 1, wherein the candidate antigens comprise a tertiary structure required for proteolytic enzyme accessibility.

13. The method of claim 1, wherein the candidate antigens comprise a defined MHCII-binding motif in the primary amino acid sequence of the predicted genes.

14. The method of claim 13, wherein the defined MHCII-binding motif is the I-$A^b$-binding motif.

15. The method of claim 1, wherein the antigenic epitopes comprise a tertiary structure required for accessibility to a MHCII binding groove.

16. The method of claim 1, wherein generating, by the processor, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network comprises:

a) encoding each amino acid of an input candidate antigen into a p-dimensional binary vector;

b) converting the binary vector to a descriptor matrix S;

c) convoluting the descriptor matrix S to a scoring matrix X; and

d) transforming the scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for a neural network.

17. The method of claim 1, wherein the antigenic epitopes are specific for an HLAtype.

18. The method of claim 3, wherein the method further comprises:

a) isolating MHCII-peptide complexes from antigen presenting cells;

b) identifying bound peptides from the MHCII-peptide complexes; and

c) training the deep neural network using the set of identified peptides.

19. The method of claim 18, wherein the antigen presenting cells are exposed to a target pathogen, commensal microorganism or diseased cell.

20. The method of claim 18, wherein the antigen presenting cells are dendritic cells.

21. The method of claim 1, wherein the candidate antigens are expressed in the pathogen, commensal microorganism or diseased cell.

22. The method of claim 1, wherein the pathogen is selected from the group consisting of a bacterium, a virus, a protozoon, and an allergen.

23. The method of claim 1, wherein the diseased cell is a cancer cell.

24. The method of claim 23, wherein the cancer cell is obtained from a subject.

25. The method of claim 24, further comprising identifying non-silent tumor specific somatic mutations from the subject specific cancer cell, wherein the ranked set of antigenic epitopes are generated from said mutations.

26. The method of claim 1, wherein the diseased cell is associated with an autoimmune disease.

27. The method of claim 22, wherein the pathogen comprises a bacteria belonging to a family selected from the group consisting of *Bacillus, Bartonella, Bordetella, Borrelia, Brucella, Campylobacter, Chlamydia and Chlamydophila, Clostridium, Corynebacterium, Enterococcus, Escherichia, Francisella, Haemophilus, Helicobacter, Legionella, Leptospira, Listeria, Mycobacterium, Mycoplasma, Neisseria, Pseudomonas, Rickettsia, Salmonella, Shigella, Staphylococcus, Streptococcus, Treponema, Ureaplasma, Vibrio,* and *Yersinia.*

28. The method of claim 1, wherein the commensal microorganism comprises a bacterium belonging to a genus selected from the group consisting of *Bacteroides, Clostridium, Faecalibacterium, Eubacterium, Ruminococcus, Peptococcus, Peptostreptococcus, Bifidobacterium, Lactobacillus a*nd *Akkermansia;* or a fungus selected from the group consisting of *Candida, Saccharomyces, Aspergillus, Penicillium, Rhodotorula, Trametes, Pleospora, Sclerotinia, Bullera,* and *Galactomyces.*

*29.* The method of claim 18, wherein the antigen presenting cell comprises a subject specific MHCII allele.

30. The method of claim 29, wherein the antigen presenting cell comprises a human MHCII allele.

31. The method of claim 1, wherein the one or more input genome sequences is obtained by whole genome or whole exome sequencing.

32. The method of claim 1, further comprising detecting whether one or more of the antigenic epitopes is present in a sample from a subject suffering from an infection, autoimmune disease, allergy, or cancer.

33. The method of claim 1, wherein the antigenic epitopes are about 9 to 20 amino acids in length.

34. The method of claim 1, wherein the deep neural network is trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 1-14979.

35. The method of claim 1, wherein the deep neural network is trained using a set of MHCII-presented peptides comprising one or more of SEQ ID NOs: 14980-15027.

36. The method of claim 1, wherein the immunological composition is a protective vaccine or tolerizing vaccine composition comprising one or more of the antigenic epitopes.

37. The method of claim 36, wherein the vaccine composition is directed to a bacterium, a virus or a protozoon.

38. The method of claim 36, wherein the vaccine composition is directed to a cancer.

39. The method of claim 36, wherein the vaccine composition is directed to an autoimmune disease or allergy, whereby administration of the vaccine induces tolerance to the one or more antigenic epitopes.

40. The method of claim 37, wherein the vaccine composition is directed to *Listeria.*

41. The method of claim 40, wherein the vaccine composition comprises one or more *Listeria* peptides selected from the peptides listed in Table 1.

42. The method of claim 41, wherein the one or more *Listeria* peptides are derived from lmo0202, lmo2558, lmo2185, or lmo0135.

43. The method of claim 42, wherein the one or more *Listeria* peptides are selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984

(APGQETQHYYGLPVADSAIDR), SEQ ID NO: 14986 (ADFRYVFDTAKATAASSYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT).

44. The method of claim 36, wherein the vaccine composition comprises autologous dendritic cells or antigen presenting cells pulsed with the one or more epitopes.

45. A method of identifying one or more immunodominant epitopes comprising:

a) expressing in a first population of immune cells one or more peptides for an immunological composition prepared according to claim 1, wherein the first population of immune cells comprise a detectable reporter gene;

b) expressing in a second population of CD4+ immune cells a TCRαβ library expressing TCRαβ pairs identified in a subject suffering from an infectious disease, autoimmunity, cancer or allergy;

c) incubating the first population of immune cells with the second population of immune cells;

d) sorting immune cells positive for the detectable reporter gene; and

e) identifying peptides bound to MHCII in immune cells positive for the detectable reporter gene.

46. The method of claim 45, wherein the TCRαβ pairs are identified by T cell profiling.

47. The method of claims 46, wherein TCRαβ pairs are determined by targeted single cell RNA-seq (TCRseq).

48. The method of claim 46, wherein T cells are analyzed by RNA-seq to determine single cells having a specific cell state and TCRαβ pairs are identified in the T cells having a specific cell state.

49. The method of claim 48, wherein the specific cell state is a pathogenic phenotype or a protective phenotype.

50. The method of claim 45, wherein the reporter is an inducible fluorescent marker protein, wherein the marker protein is induced when a TCR$\alpha\beta$ pair detects an MHCII epitope.

51. The method of claim 50, wherein the fluorescent marker is under control of the IL-2 promoter.

52. The method of claims 45, further comprising formulating a vaccine composition comprising one or more of the immunodominant epitopes.

53. The method of claim 52, wherein the vaccine is a protective vaccine or a tolerizing vaccine.

54. The method of claim 52, wherein the vaccine comprises autologous dendritic ceils or antigen presenting cells pulsed with one or more of the immunodominant epitopes.

55. A method of identifying peptide antigens from a live bacterial pathogen comprising:
isolating MHCII-peptide complexes from dendritic cells exposed to a bacterial pathogen;
isolating the peptides from the MHCII-peptide complexes; and
sequencing the isolated peptides.

56. A method of determining immune related health status in a subject comprising:

a) preparing one or more peptides for an immunological composition according to claim 1;

b) exposing the one or more peptides to immune cells obtained from the subject; and

c) measuring cytokine secretion.

57. The method of claim 56, wherein the cytokines measured comprise IL-2, IFN-$\gamma$, IL-17, and/or IL-10.

58. The method of claim 56, further comprising measuring immune ceil types reactive to the epitopes in the subject.

59. The method of claim 58, wherein the immune cell types comprise Treg, Thl, Thl7 and/or Th2 cells.

60. The method of claim 58, wherein the immune cells are measured using MHCII tetramers.

61. A method of determining immune related health status in a subject comprising:

a) preparing one or more peptides for an immunological composition according to claim 1; and

b) measuring immune cell types reactive to the epitopes in the subject.

62. The method of claim 61, wherein the immune cell types comprise Treg, Thl, Thl7 and/or Th2 cells.

63. The method of claim 61, wherein the immune cells are measured using MHCII tetramers.

64. A method of constructing a deep neural network for identifying MHCII antigenic epitopes comprising:

a) isolating MHCII-peptide complexes from antigen presenting cells expressing a MHCII type;

b) isolating peptides from the MHCII-peptide complexes;

c) sequencing the isolated peptides; and

d) training a deep neural network for predicting antigenic epitopes for the MHCII type using the set of isolated peptides.

65. The method of claim 64, wherein the deep neural network is trained by testing randomly initialized parameters for the peptides.

66. The method of claim 65, wherein the parameters are tested in a three-fold cross validation scheme.

67. The method of claim 65, wherein the parameters comprise cellular localization, inter-domain structure, domain size and/or tertiary structure.

68. The method of claim 67, wherein cellular localization comprises peptides derived from extracellular, intracellular or transmembrane proteins.

69. A peptide set for training a MHCII neural network comprising one or more of SEQ ID NOs: 1-14979.

70. A peptide set for training a MHCII neural network comprising one or more of SEQ ID Nos: 14980-15027.

71. A *Listeria* vaccine comprising one or more peptides selected from the peptides listed in Table 1.

72. The *Listeria* vaccine of claim 71, wherein the one or more peptides is derived from lmo0202, lmo2558, lmo2185, or lmo0135.

73. The *Listeria* vaccine of claim 72, wherein the one or more *Listeria* peptides is selected from the group consisting of SEQ ID NO: 14981 (WNEKYAQAYPNVSAKI), SEQ ID NO: 14984 (APGQETQHYYGLPVADSAIDR), SEQ ID NO: 14986 (ADFRYVFDTAKATAASSYPG), and SEQ ID NO: 14997 (VDDTTVKFTLPTVAPAFENT).

74. The *Listeria* vaccine of claim 71, wherein the vaccine comprises autologous dendritic cells or antigen presenting cells pulsed with the one or more peptides.

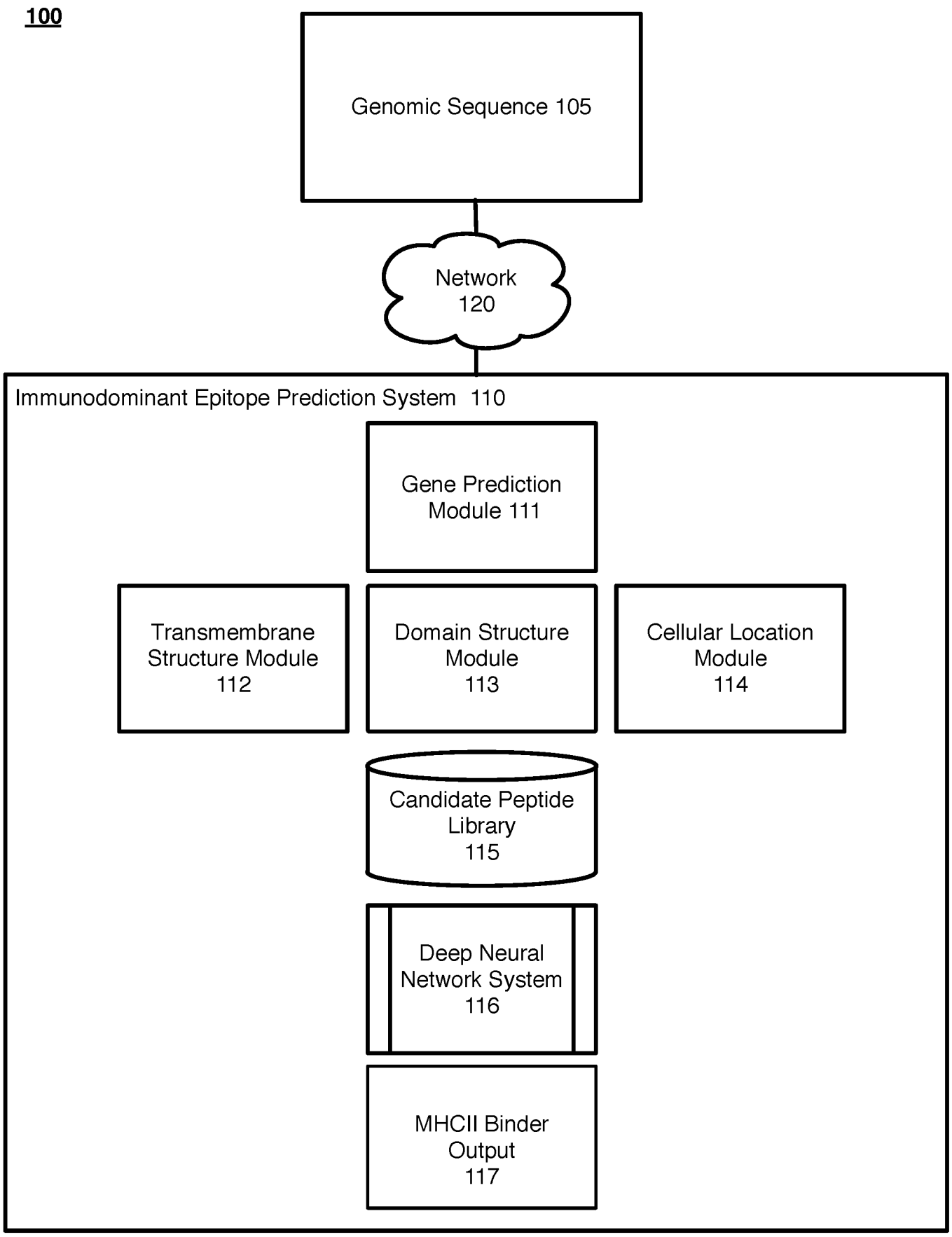

100
Genomic Sequence 105
Network
120
Immunodominant Epitope Prediction System 110
Gene Prediction
Module 111
Transmembrane
Structure Module
112
Domain Structure
Module
113
Cellular Location
Module
114
Candidate Peptide
Library
115
Deep Neural
Network System
116
MHCII Binder
Output
117

FIG. 1

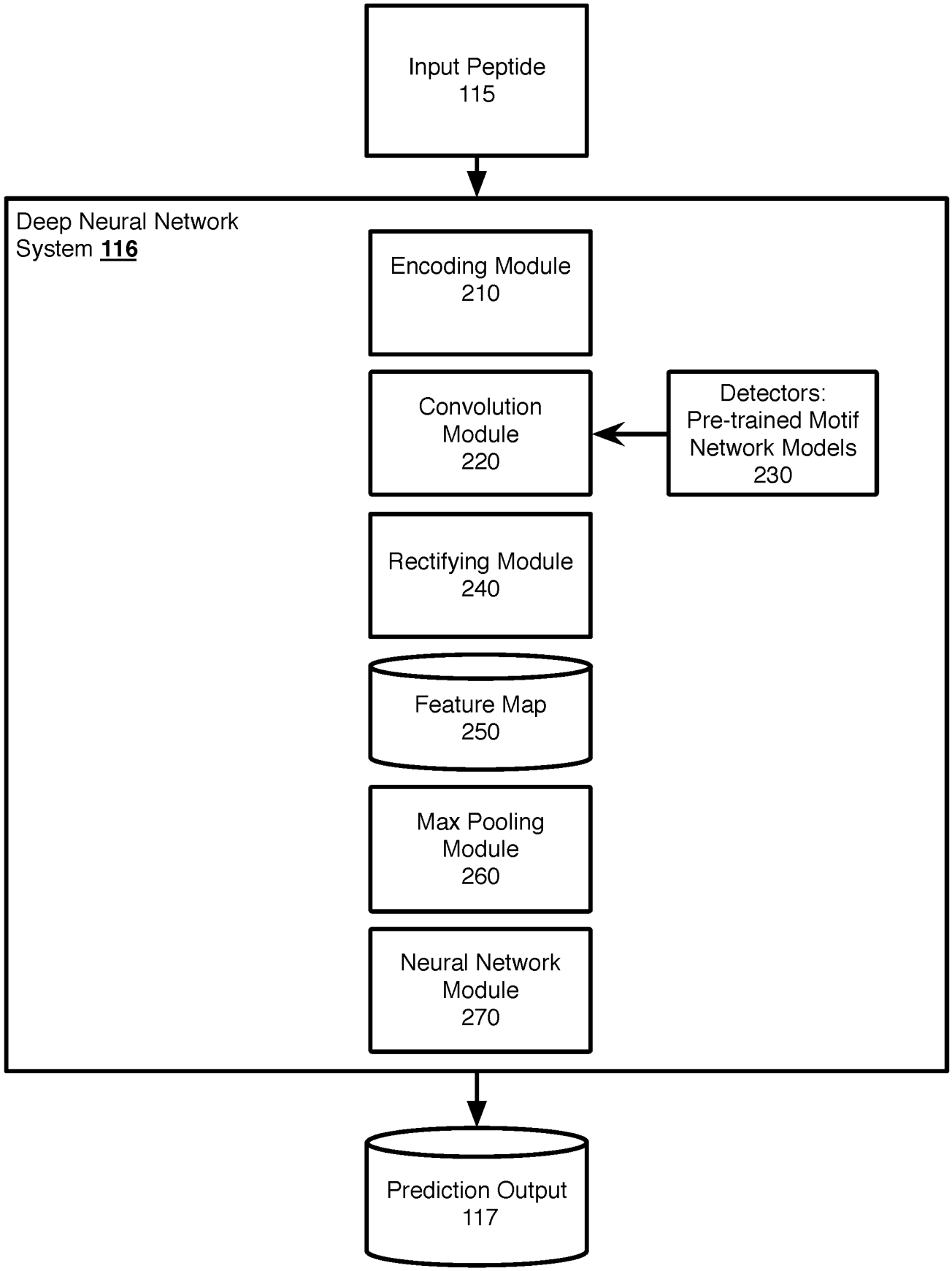

Input Peptide
115
Deep Neural Network
System 116
Encoding Module
210
Convolution
Module
220
Detectors:
Pre-trained Motif
Network Models
230
Rectifying Module
240
Feature Map
250
Max Pooling
Module
260
Neural Network
Module
270
Prediction Output
117

FIG. 2

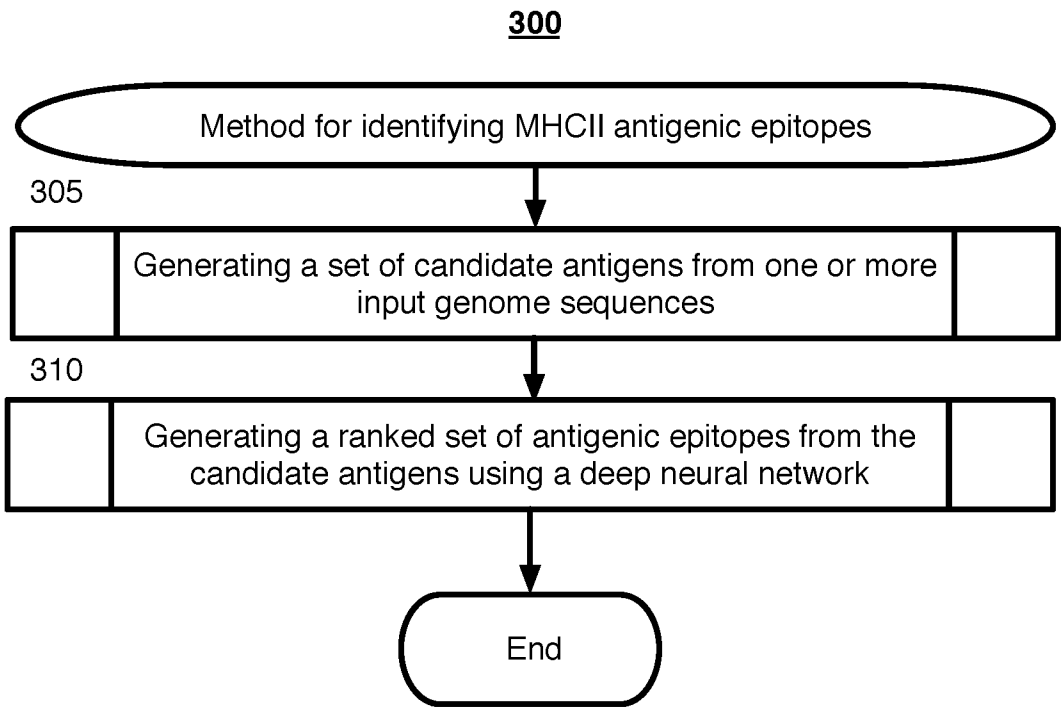

300
Method for identifying MHCII antigenic epitopes
305
Generating a set of candidate antigens from one or more input genome sequences
310
Generating a ranked set of antigenic epitopes from the candidate antigens using a deep neural network
End

FIG. 3

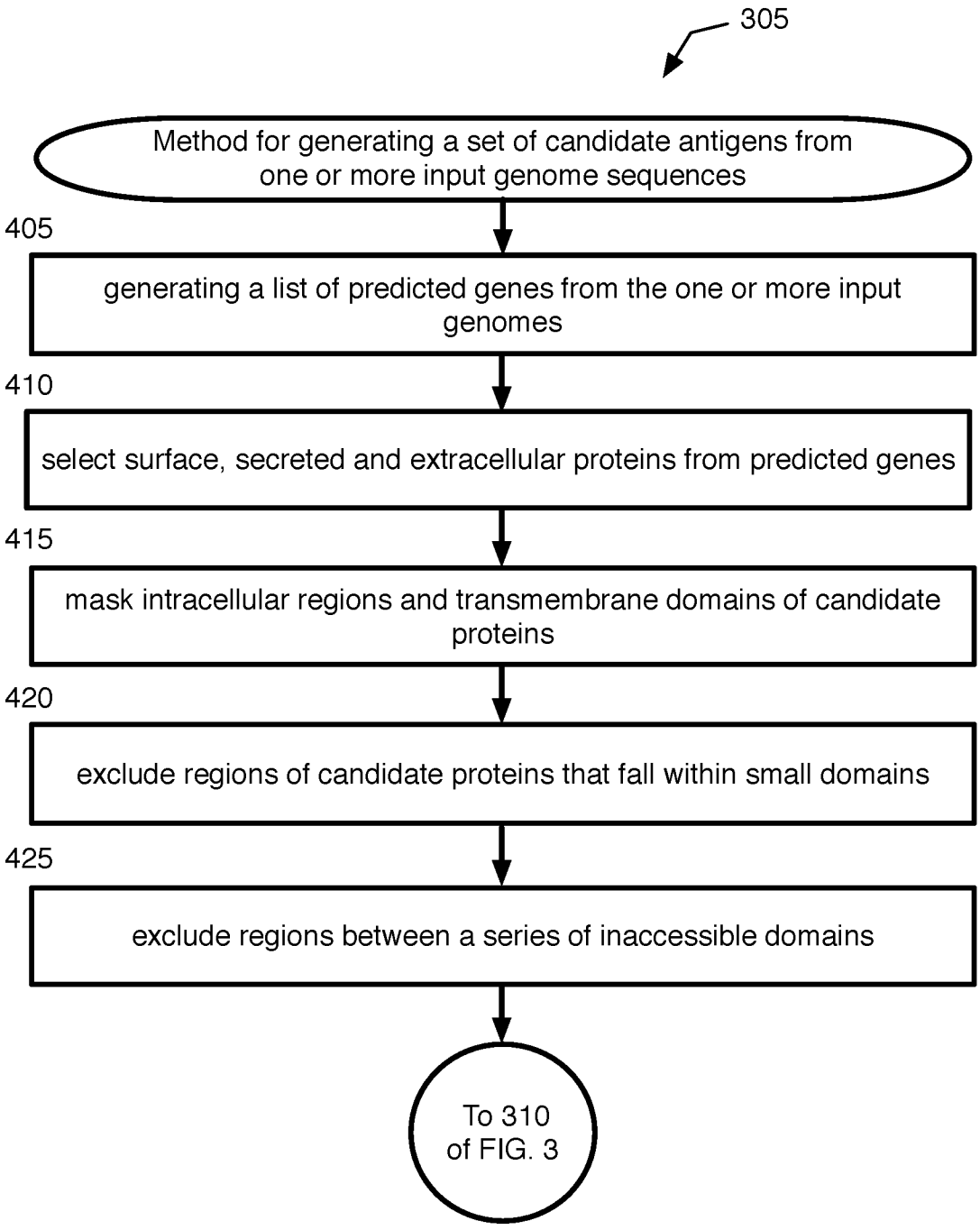

305
Method for generating a set of candidate antigens from one or more input genome sequences
405
generating a list of predicted genes from the one or more input genomes
410
select surface, secreted and extracellular proteins from predicted genes
415
mask intracellular regions and transmembrane domains of candidate proteins
420
exclude regions of candidate proteins that fall within small domains
425
exclude regions between a series of inaccessible domains
To 310 of FIG. 3

FIG. 4

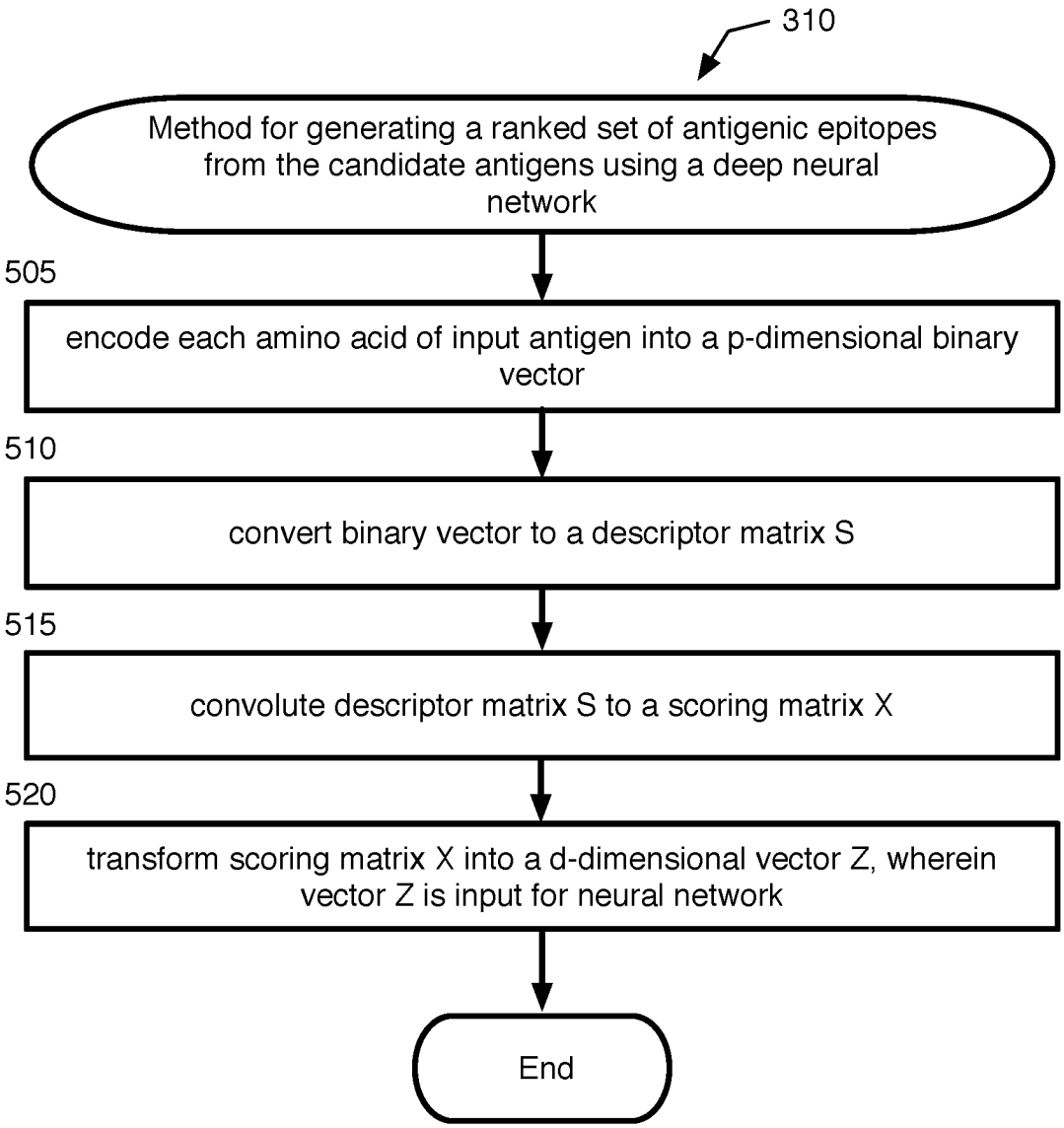

310
Method for generating a ranked set of antigenic epitopes from the candidate antigens using a deep neural network
505
encode each amino acid of input antigen into a p-dimensional binary vector
510
convert binary vector to a descriptor matrix S
515
convolute descriptor matrix S to a scoring matrix X
520
transform scoring matrix X into a d-dimensional vector Z, wherein vector Z is input for neural network
End

FIG. 5

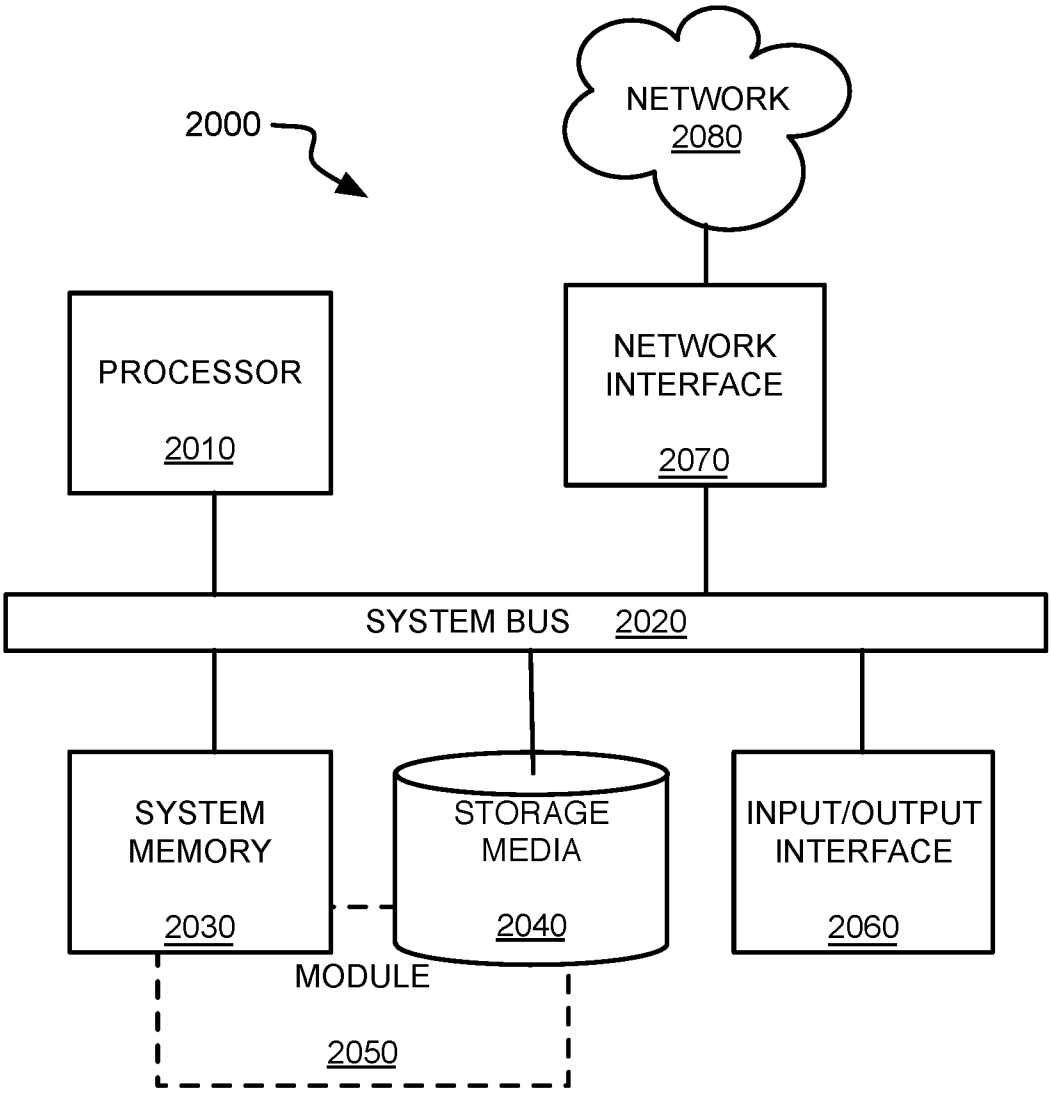

2000
NETWORK
2080
PROCESSOR
2010
NETWORK
INTERFACE
2070
SYSTEM BUS 2020
SYSTEM
MEMORY
2030
STORAGE
MEDIA
2040
INPUT/OUTPUT
INTERFACE
2060
MODULE
2050

FIG. 6

WO 2019/006022

PCT/US2018/039843

FIG. 7A

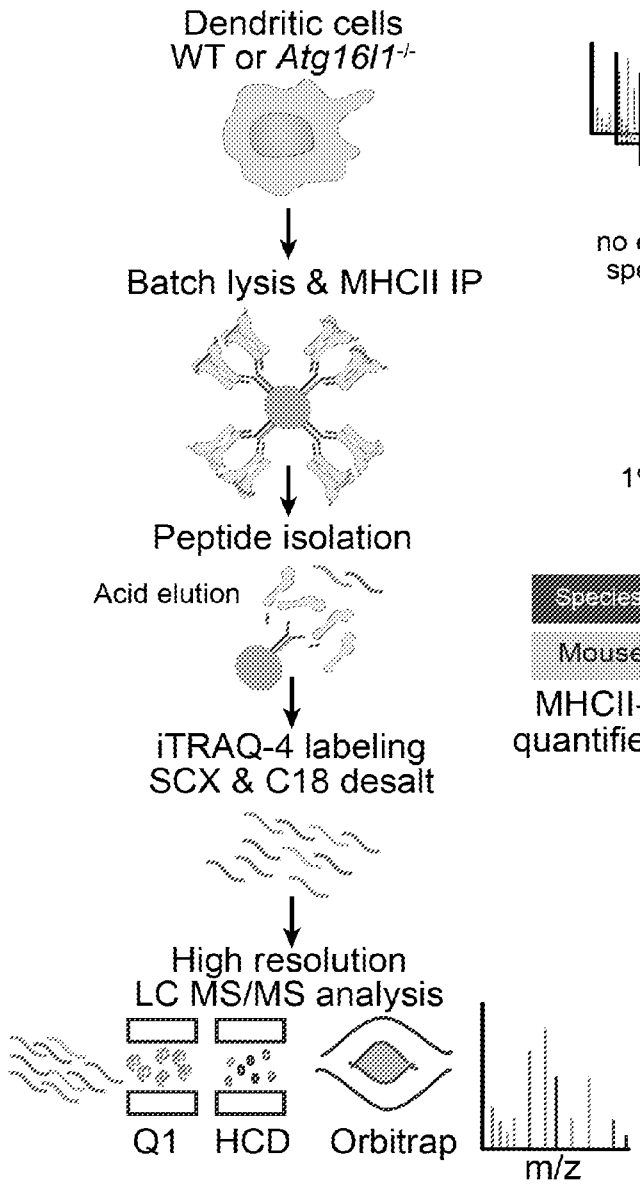



FIG. 7B

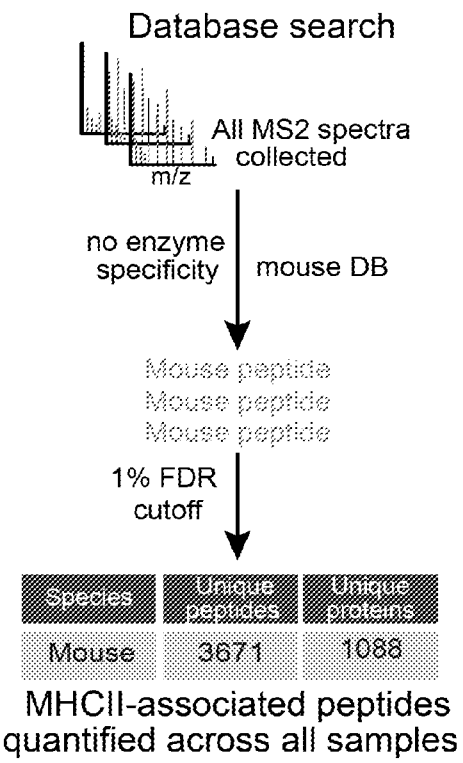



FIG. 7C

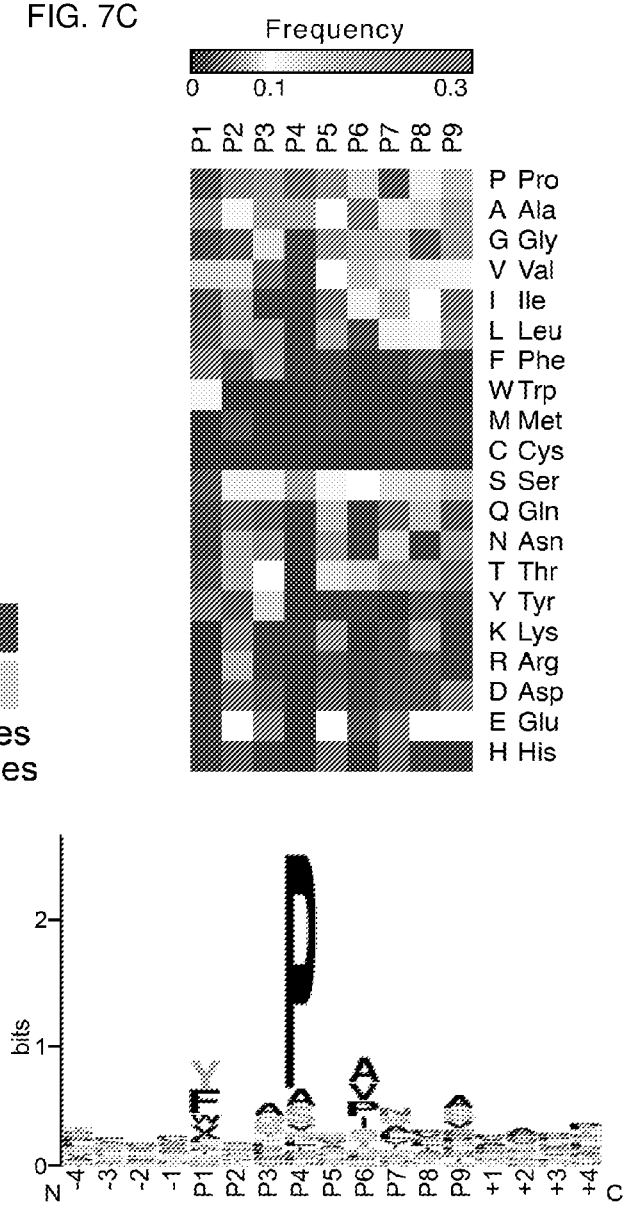


FIG. 8A

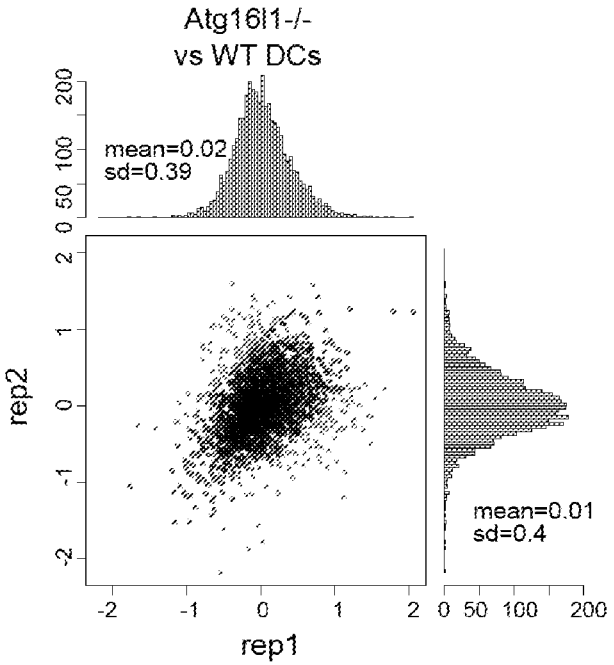



FIG. 8B

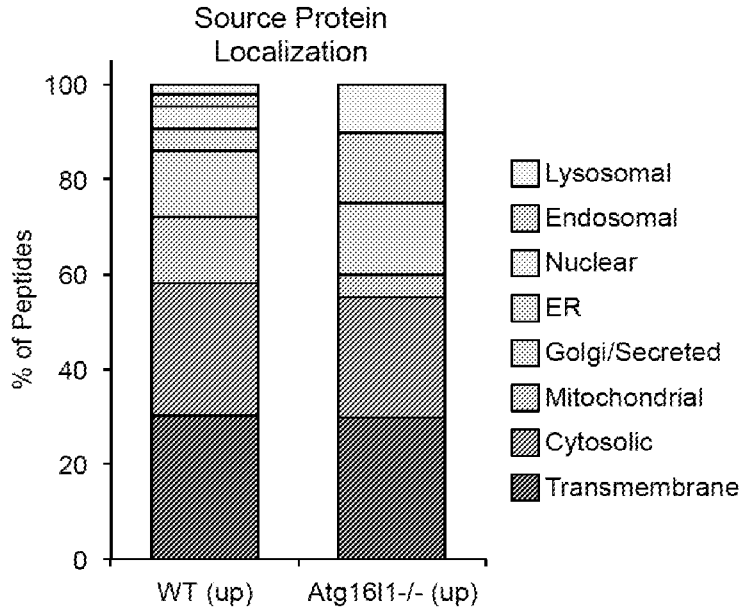
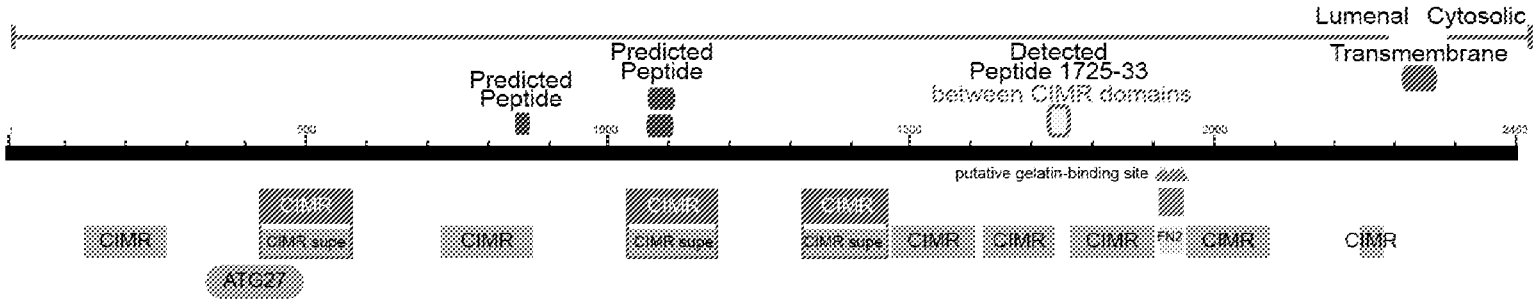



FIG. 8C

Igf2r (Cation-Independent Mannose-6 Phosphate Receptor): Endolysosomal




| Name | Accession | Description |
|---|---|---|
| FN2 | cd00062 | Fibronectin Type II domain |
| CIMR | pfam00878 | Cation-independent mannose-6-phosphate receptor repeat |
| CIMR super family | cl03002 | Cation-independent mannose-6-phosphate receptor repeat |
| ATG27 | pfam09451 | Autophagy-related protein 27 |

WO 2019/006022
PCT/US2018/039843

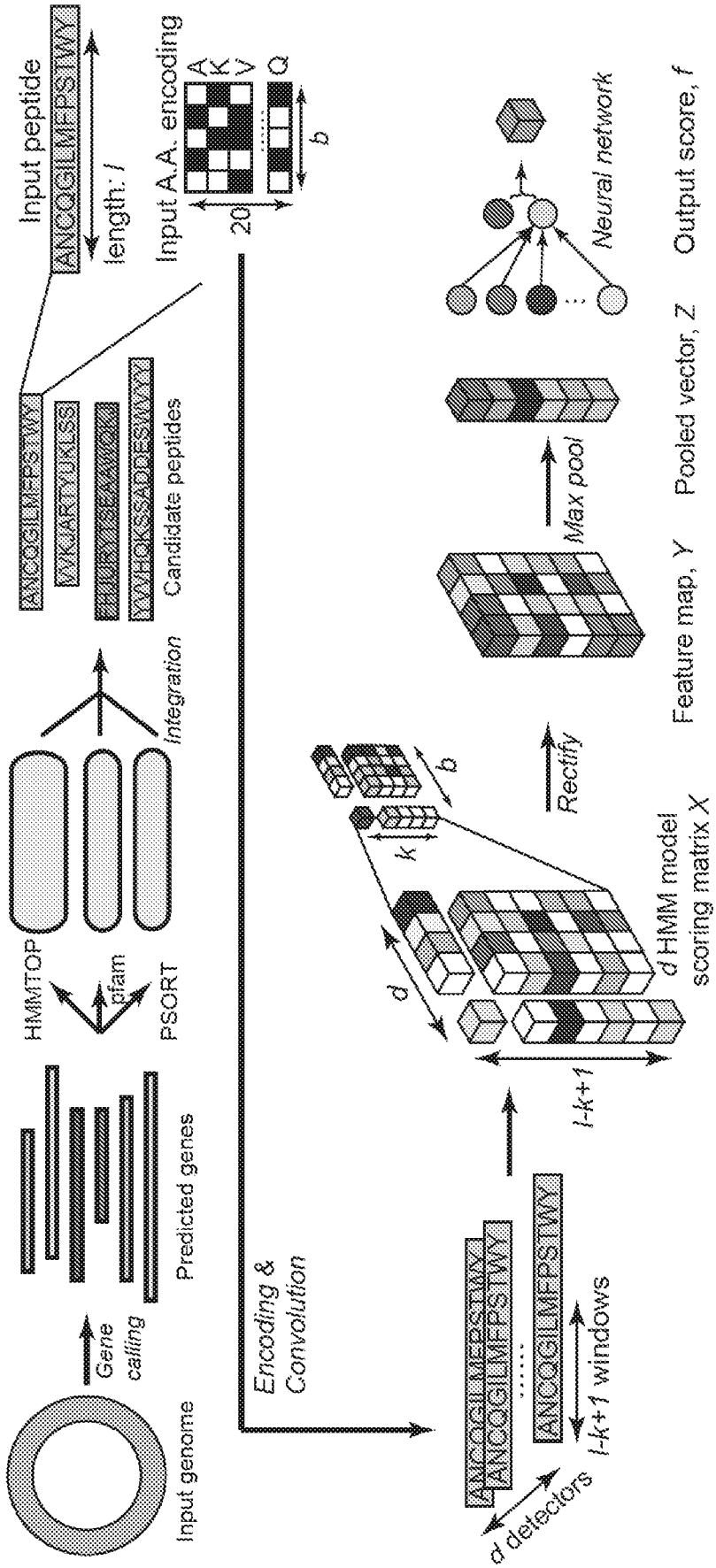

Input genome
Gene calling
Predicted genes
HMMTOP
pfam
PSORT
Integration
Candidate peptides
Input peptide
length: l
Input A.A. encoding
20
b
Encoding & Convolution
d detectors
l-k+1 windows
d HMM model scoring matrix X
k
b
d
l-k+1
Rectify
Feature map, Y
Max pool
Pooled vector, Z
Neural network
Output score, f



FIG. 8D


WT
Atg16l1-/-

| Id | Description |
|---|---|
| SMFLHINETYLLMEQ | TBC1 domain family member 8B |
| RFTYQNAAAQPETGPG | Tumor necrosis factor receptor superfamily member 10B |
| SKTVYALPTIAFAF | Sodium-coupled neutral amino acid transporter 1 |
| SPRDDARVYHDPSTQEPVT | Plectin |
| FQHPYNTLKYPnGEGG | Kremen protein 1 |
| EKWEETELDQLFSGPLPGRLP | Tyrosine-protein kinase Mer |
| DVVDRVLKSPVASIVYEL | "Asparagine--tRNA ligase, cytoplasmic" |
| DRRYNPPIIHQGFQ | C-C motif chemokine 6 |
| RTDWEAPTGQPSAPG | TBC1 domain family member 1 |
| RPGLSYHYAHSHLAEEEG | Zinc finger protein ubi-d4 |
| NHTLFHNPHVNPLPT | "Cytochrome c oxidase subunit 6A1, mitochondrial" |
| ETTVNPSGKNVSYNISSTVRVVLnSMDVHS | Tyrosine-protein phosphatase non-receptor type substrate 1 |
| DQKFYNFVVIHARAD | TIR domain-containing adapter molecule 1 |
| EKTGQEYKPGNPSAAAVQ | Bifunctional glutamate/proline--tRNA ligase |
| SKQEYDESGPSMHR | "Actin, cytoplasmic 1" |
| LRDNYRFAHTNIESLV | Protein disulfide-isomerase A3 |
| FSFGKINPHYTVKGLE | Protein Olfr731 |
| DGIHHWTPPKGHVDPGEN | Bis(5'-nucleosyl)-tetraphosphatase [asymmetrical] |
| GNHTLFHNPHVNPLPTG | "Cytochrome c oxidase subunit 6A1, mitochondrial" |
| GDIEHPRAIALDPR | Prolow-density lipoprotein receptor-related protein 1 |
| NLKHVAGAAAAGAVV | Major prion protein |
| DGREVSSVEAPSVSPARHQ | Low-density lipoprotein receptor-related protein 12 |
| KGQDHVDFSQPEPHT | Semaphorin-7A |
| RYMAPEVLESR | TGF-beta receptor type-2 |
| IEVAPPKAHEVR | Alcohol dehydrogenase 1 |
| RKIVYEGPELNHAF | Prolow-density lipoprotein receptor-related protein 1 |
| RQTTQYPEANAL | Complement C1q subcomponent subunit C |
| EAIETIVAAMSNLVPPVE | Guanine nucleotide-binding protein G(s) subunit alpha isoforms XLas |
| EDRRYNPPIIHQ | C-C motif chemokine 6 |
| GIKADKADKAAASGPASASAP | Integral membrane protein 2C |
| FVnYSNAVTLHLGNn | SLIT and NTRK-like protein 2 |
| EmEKEDRRYNPPIIHQGFQ | C-C motif chemokine 6 |
| EPRVLPYHPAQPGQAT | Chondroitin sulfate synthase 2 |
| GRRTDWEAPTGQPSAPGP | TBC1 domain family member 1 |
| DKNGNPLNSVFVAPAVTPVKGVL | Acid sphingomyelinase-like phosphodiesterase 3a |
| GDLEYQMSTTARAKR | Junction plakoglobin |
| ATPRDYYFALAHTVRDH | "Glycogen phosphorylase, liver form" |
| SSIASFKDPKISTSLPV | Plastin-2 |
| LPQDPTFHTPLDLYAYAR | Galectin-3-binding protein |
| DVDFDFAGPAIHGSAV | Voltage-dependent anion-selective channel protein 2 |
| GYPTAYPAAAPAYNPS | Protein FAM168A |
| DSPDVVYSPETIEARYE | Inositol-3-phosphate synthase 1 |
| TPDPNTAYASYPSASVEN | Synaptogyrin-2 |
| ESPDQWSSSSPHSASDW | Neurogenic locus notch homolog protein 2 |
| VDFDFAGPAIHGSAV | Voltage-dependent anion-selective channel protein 2 |
| DPTPDPNTAYASYPSASVEN | Synaptogyrin-2 |
| KLLEAKGPVTDVAYSHD | WD repeat-containing protein 1 |
| APVEGASPTSISAVDG | "Class II histocompatibility antigen, M alpha chain" |
| KKELEEQLGPVAEETR | Apolipoprotein E |
| LVEKIF | BPI fold-containing family B member 4 |
| SSVAFIKPGVNKAEVT | SLAM family member 5 |
| SPDQWSSSSPHSASDW | Neurogenic locus notch homolog protein 2 |
| APETRSQVTPSVPSEALDD | Low-density lipoprotein receptor-related protein 10 |
| YnIPFTEKnILYSAFLWRT | Protein Vmn2r60 |
| SPESPDQWSSSSPHSASDW | Neurogenic locus notch homolog protein 2 |
| TPSPESPDQWSSSSPHSASDW | Neurogenic locus notch homolog protein 2 |
| DGDSYRSPWSNKYDPP | F-actin-capping protein subunit beta |
| IGNERFRCPEAIFQPS | Beta-actin-like protein 2 |
| DGDSYRSPWSNKYDPPL | F-actin-capping protein subunit beta |
| RKIVYEGPELNHAFG | Prolow-density lipoprotein receptor-related protein 1 |

FIG. 8E

FIG. 9A

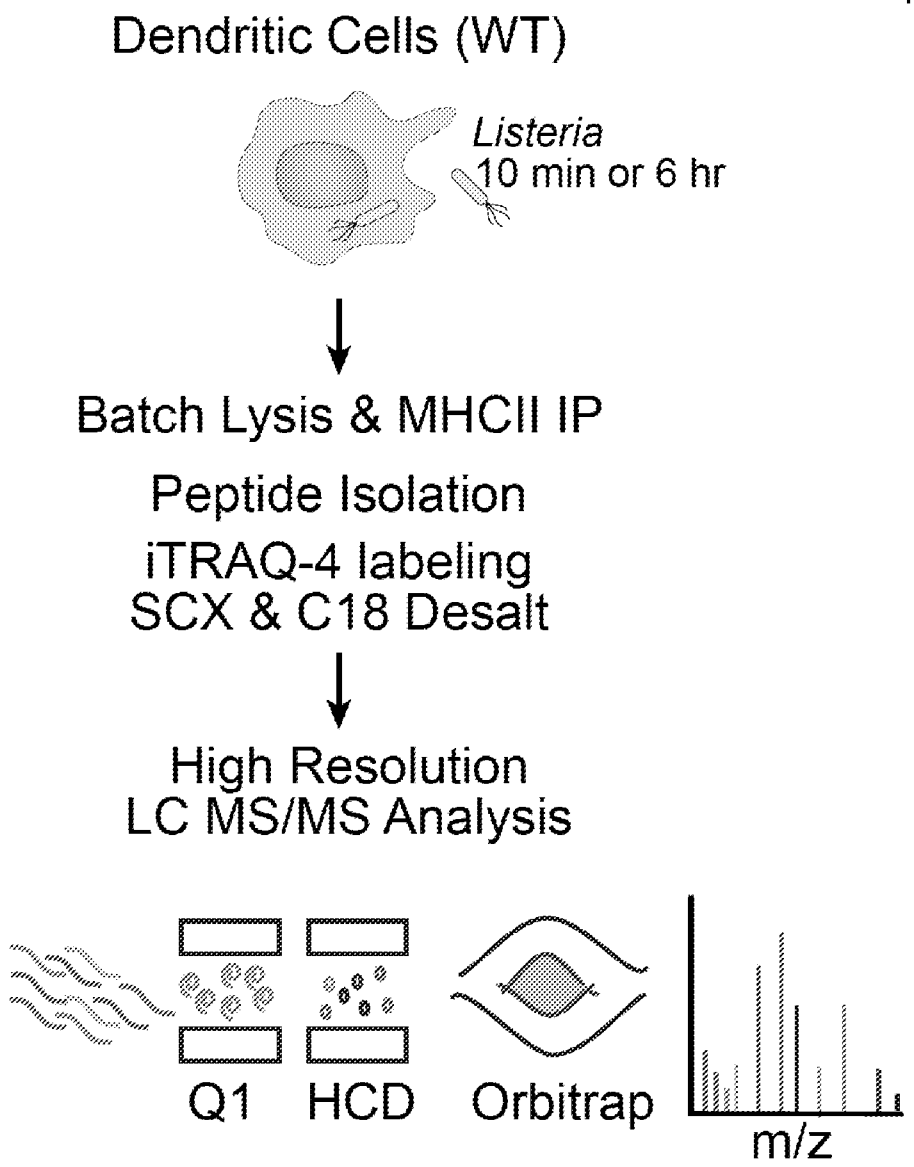
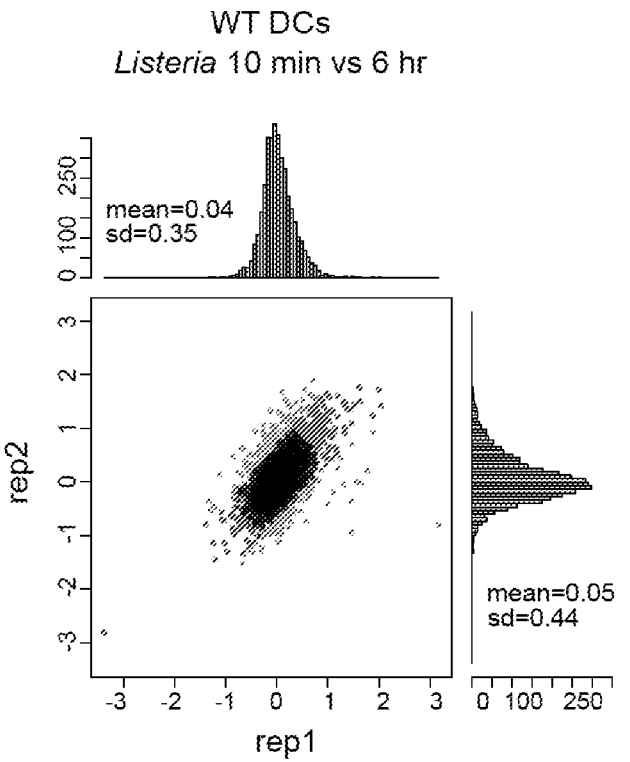



FIG. 9B




| Species | Unique Peptides | Unique Proteins |
|---|---|---|
| Listeria | 48 | 21 |

WO 2019/006022
PCT/US2018/039843

FIG. 9C

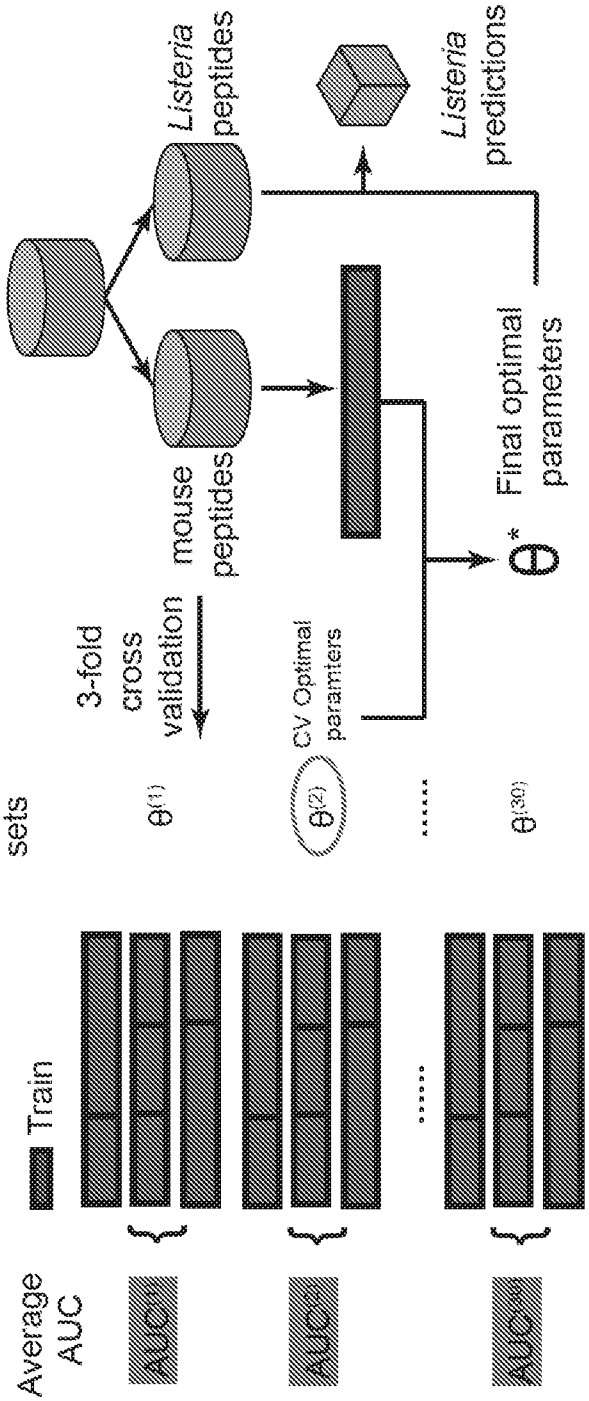



FIG. 9D

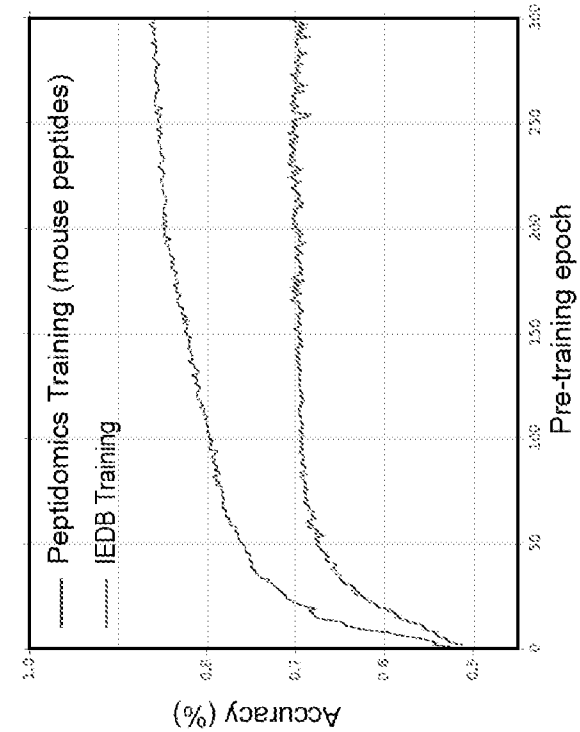
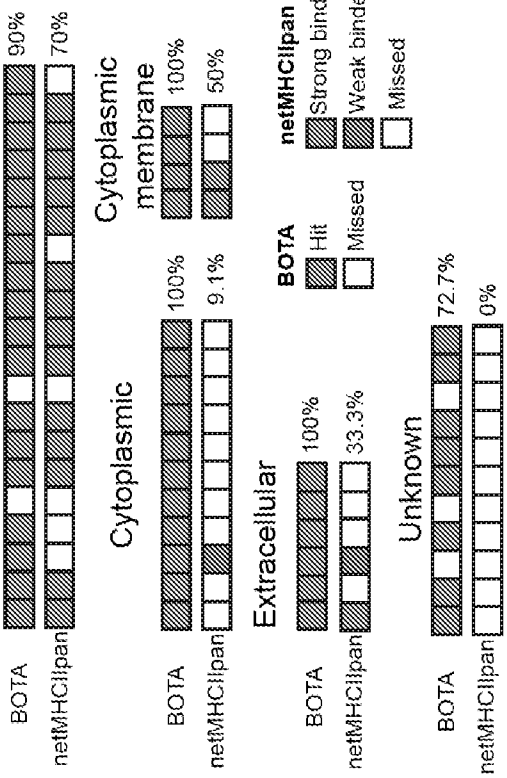
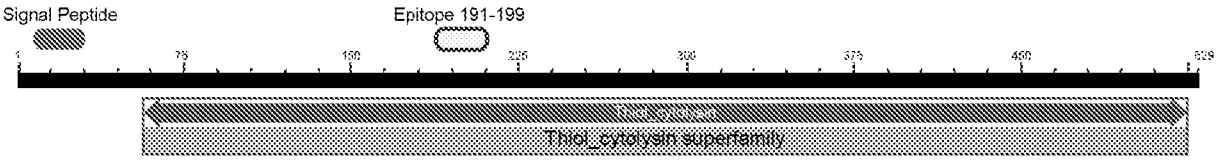



FIG. 9E

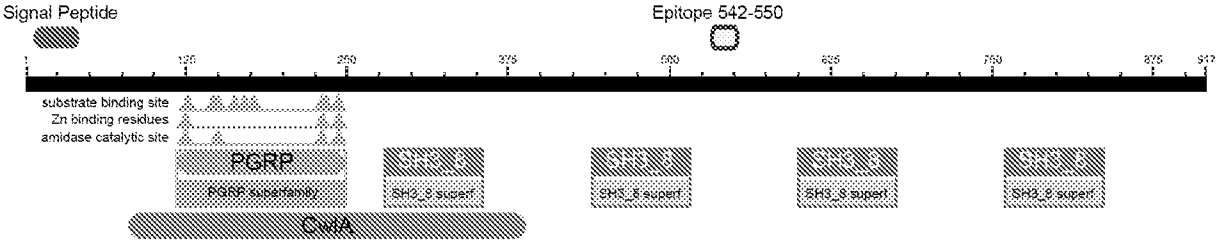


lmo0202 (hly) listeriolysin O precursor: Secreted



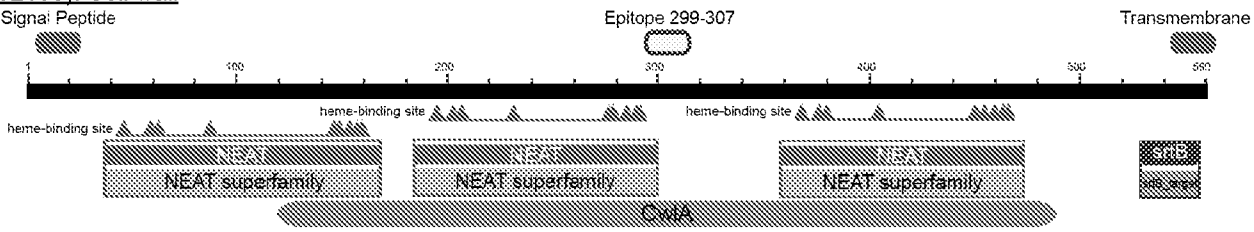

| Name | Accession | Description |
|---|---|---|
| Thiol_cytolysin | pfam01289 | Thiol-activated cytolysin |

lmo2558 (ami) autolysin, amidase: Secreted



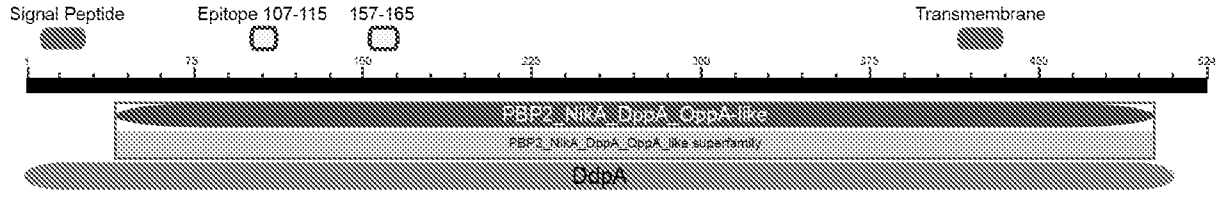

| Name | Accession | Description |
|---|---|---|
| PGRP | cd06583 | Peptidoglycan recognition proteins (PGRPs), pattern recognition receptors |
| SH3_8 | pfam13457 | SH3-like domain |
| CwlA | COG5632 | N-acetylmuramoyl-L-alanine amidase CwlA [Cell wall/membrane/envelope biogenesis] |

lmo2185 (lmo2185): Cell wall




| Name | Accession | Description |
|---|---|---|
| NEAT | cd06920 | NEAr Transport domain, a component of cell surface proteins |
| srtB_target | TIGR03063 | sortase B cell surface sorting signal |
| NEAT | COG5386 | Heme-binding NEAT domain [Inorganic ion transport and metabolism] |

lmo0135 peptide ABC transporter substrate-binding protein: Cell wall




| Name | Accession | Description |
|---|---|---|
| PBP2_NikA_DppA_OppA_like | cd00995 | ABC-type nickel/oligopeptide-like import system |
| DdpA | COG0747 | ABC-type transport system, periplasmic component [Amino acid transport and metabolism] |

FIG. 10A

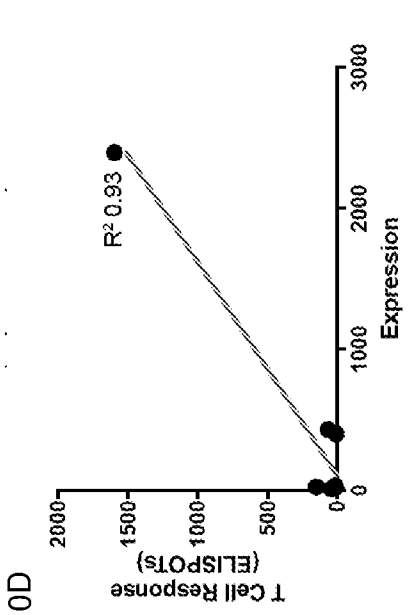



FIG. 10D

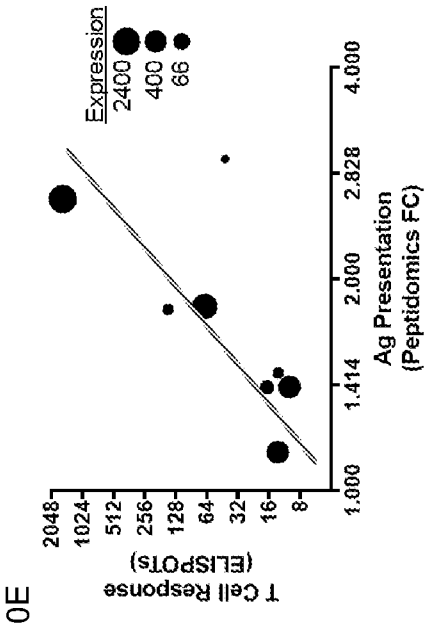



FIG. 10E

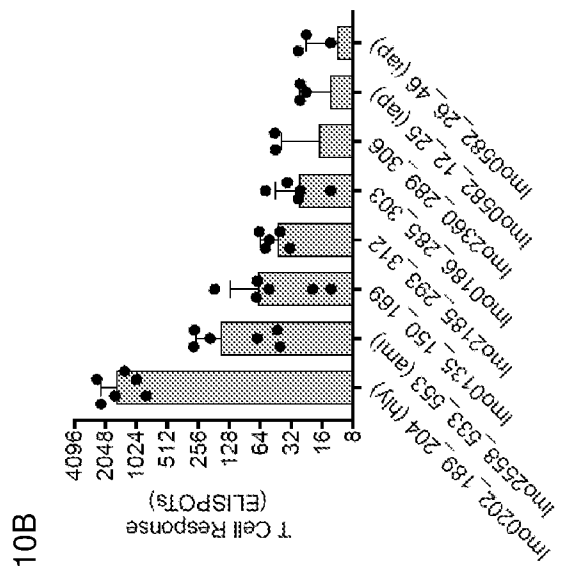



FIG. 10B




FIG. 10C

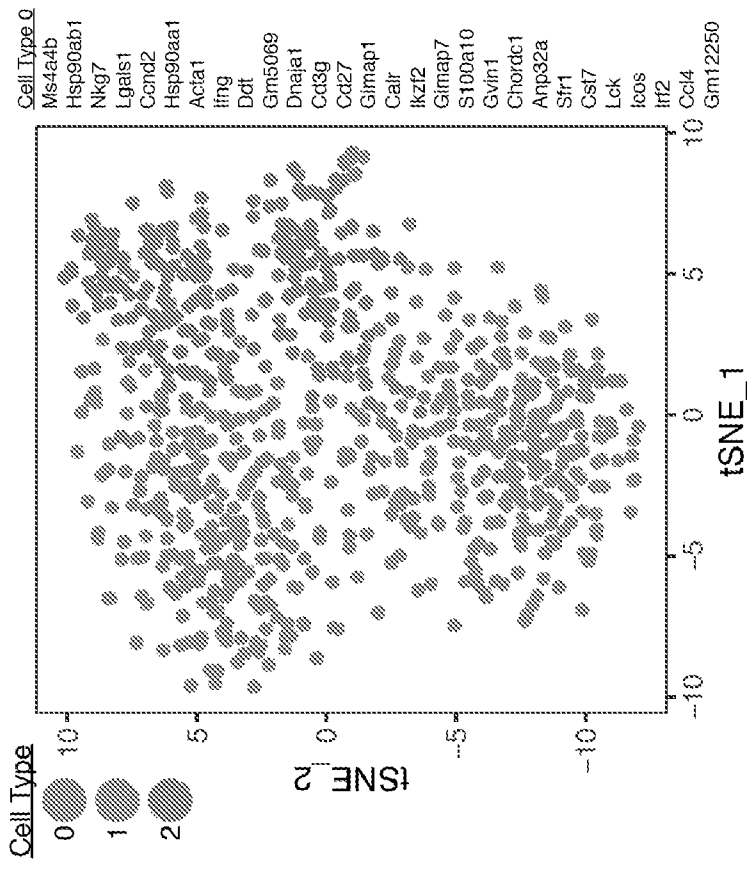



FIG. 11A

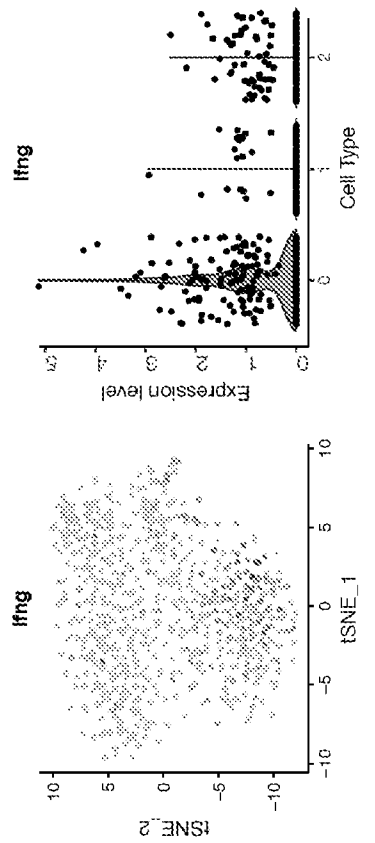



FIG. 11B

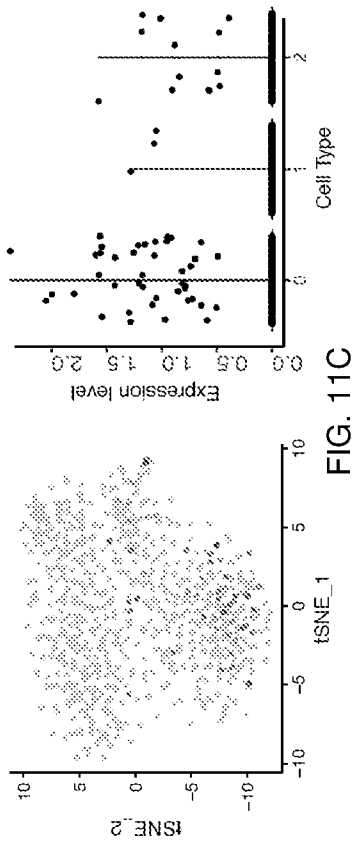
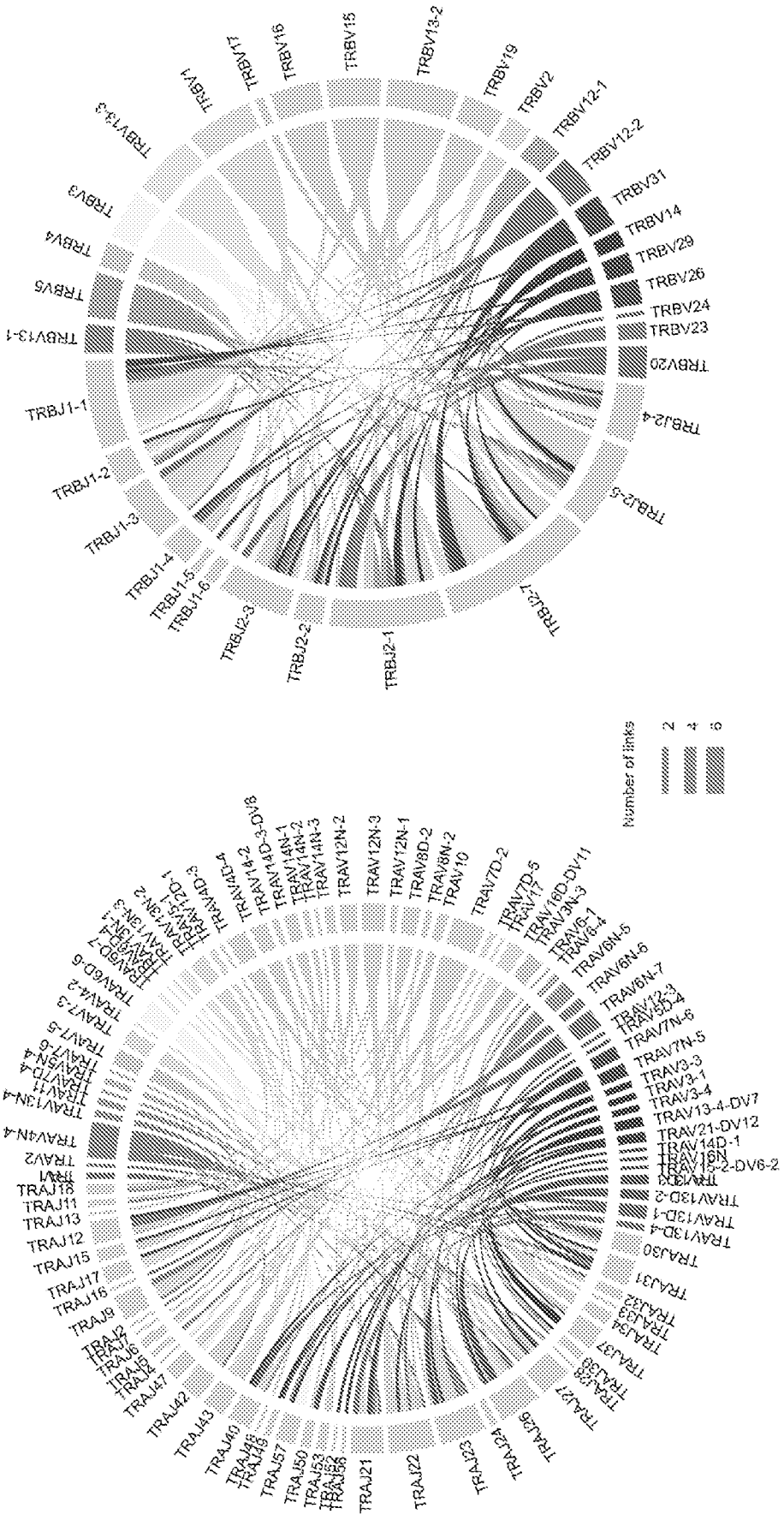



FIG. 11C


TRBV15
TRBV13-2
TRBV19
TRBV2
TRBV12-1
TRBV12-2
TRBV31
TRBV14
TRBV29
TRBV26
TRBV24
TRBV23
TRBV20
TRBJ2-4
TRBJ2-5
TRBJ2-7
TRBJ2-1
TRBJ2-2
TRBJ2-3
TRBJ1-6
TRBJ1-5
TRBJ1-4
TRBJ1-3
TRBJ1-2
TRBJ1-1
TRBV13-1
TRBV5
TRBV4
TRBV3
TRBV13-3
TRBV1
TRBV17
TRBV16
Number of links
2
4
6
TRAV12N-3
TRAV12N-1
TRAV8D-2
TRAV6N-2
TRAV10
TRAV7D-2
TRAV7D-5
TRAV17
TRAV16D-DV11
TRAV3N-3
TRAV6-1
TRAV6-4
TRAV6N-5
TRAV6N-6
TRAV6N-7
TRAV12-3
TRAV6D-4
TRAV7N-6
TRAV7N-5
TRAV3-3
TRAV3-1
TRAV3-4
TRAV13-4-DV7
TRAV21-DV12
TRAV14D-1
TRAV16N
TRAV13-2-DV6-2
TRAV13D-2
TRAV13D-1
TRAV13D-4
TRAJ30
TRAJ31
TRAJ32
TRAJ33
TRAJ34
TRAJ37
TRAJ38
TRAJ21
TRAJ22
TRAJ23
TRAV24
TRAJ39
TRAJ42
TRAJ43
TRAJ40
TRAJ48
TRAJ49
TRAJ57
TRAJ50
TRAJ53
TRAJ54
TRAJ56
TRAJ52
TRAJ58
TRAJ47
TRAJ45
TRAJ6
TRAJ7
TRAJ5
TRAJ4
TRAJ9
TRAJ16
TRAJ17
TRAJ15
TRAJ12
TRAJ13
TRAJ11
TRAJ18
TRAJ14
TRAV2
TRAV4N-4
TRAV13N-4
TRAV5N-4
TRAV13-1
TRAV13D-4
TRAV5-6
TRAV1-5
TRAV7-3
TRAV4-2
TRAV6D-6
TRAV6D-7
TRAV5N-4
TRAV13N-3
TRAV13N-2
TRAV5-1-DV5-1
TRAV4D-3
TRAV4D-4
TRAV14-2
TRAV14D-3-DV8
TRAV14N-1
TRAV14N-2
TRAV14N-3
TRAV12N-2

16/28

FIG. 11D

FIG. 12A

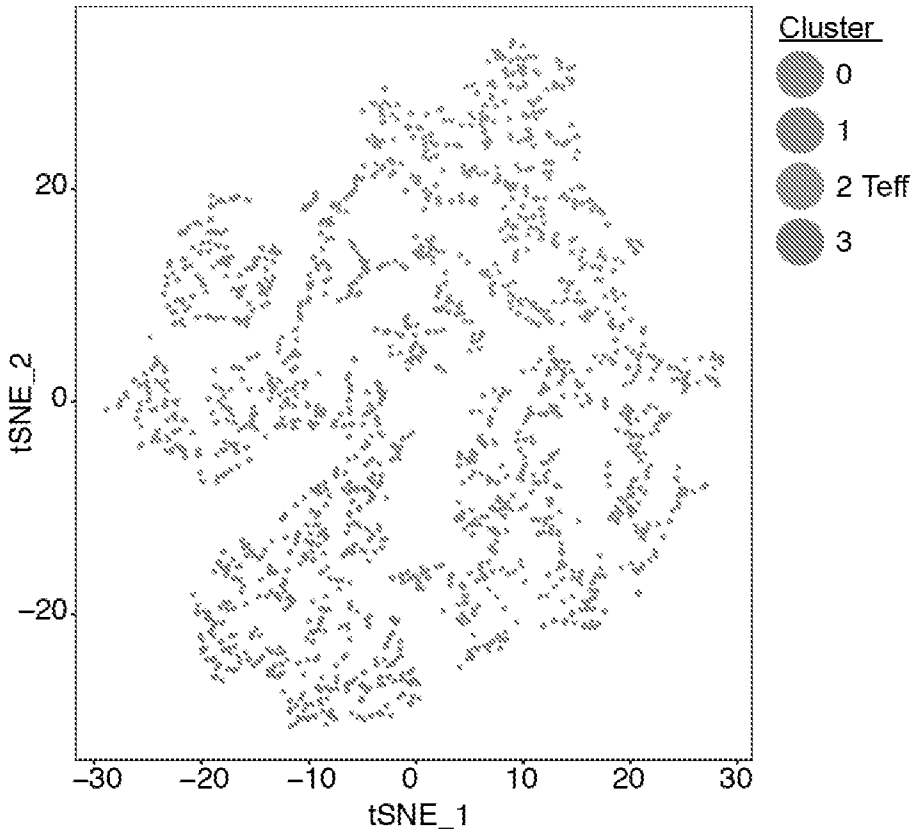
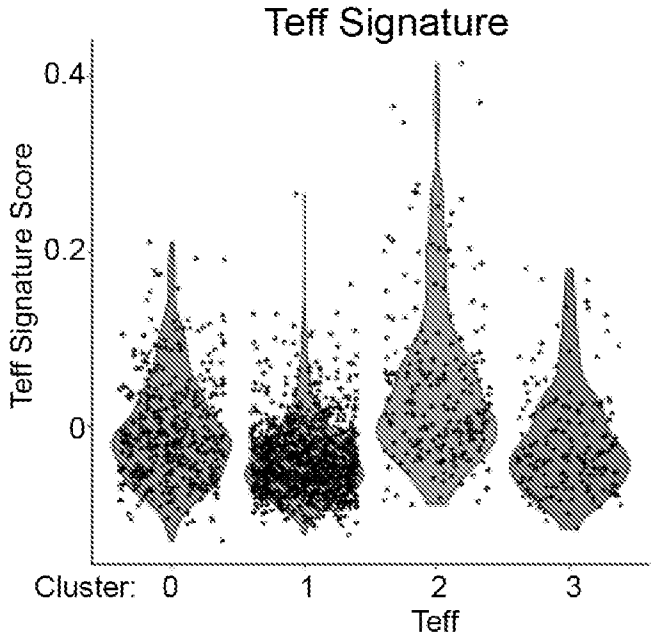



FIG. 12B



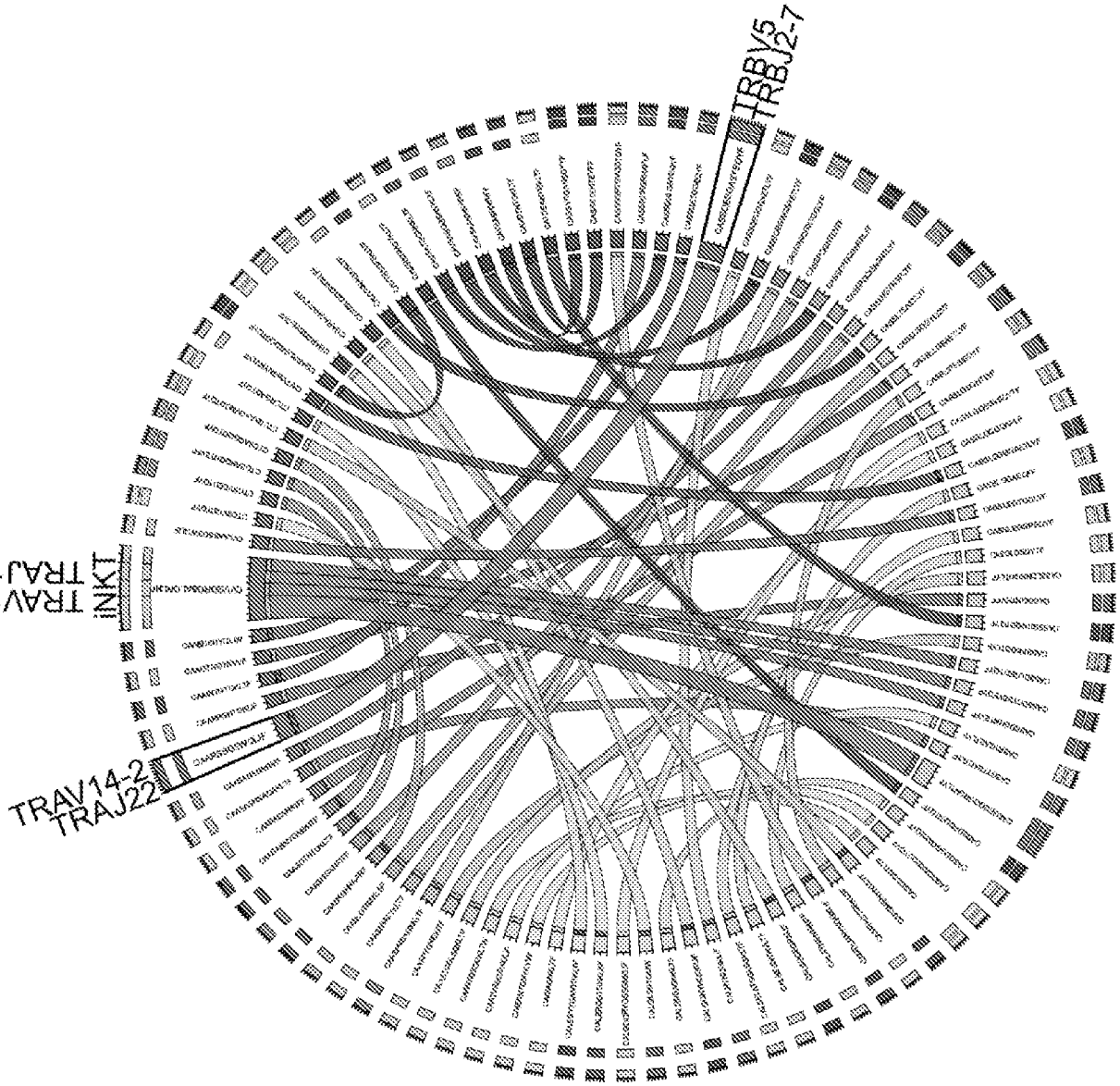





FIG. 12C

FIG. 13A

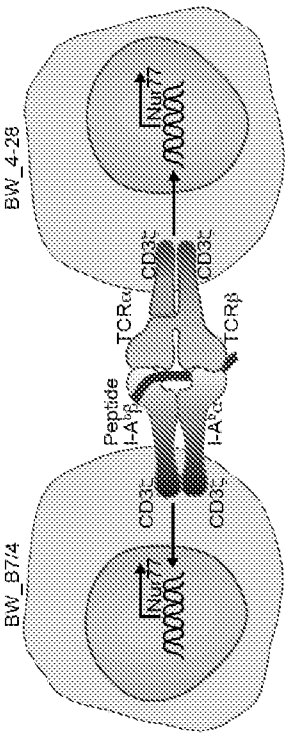



FIG. 13B

FIG. 14A

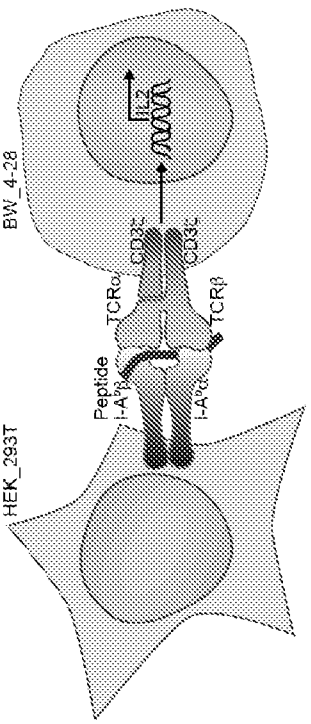



FIG. 14B

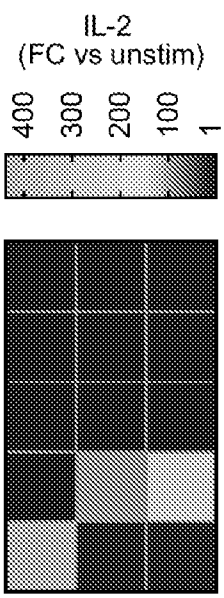



FIG. 14C

WO 2019/006022

FIG. 15A

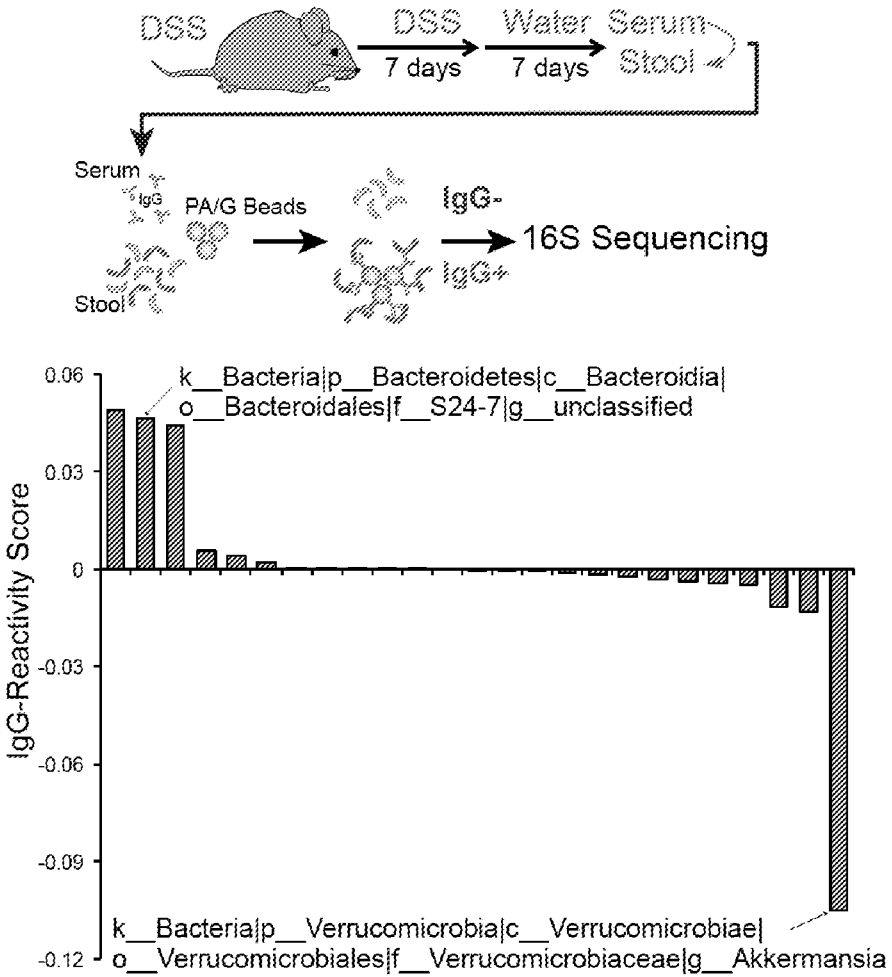
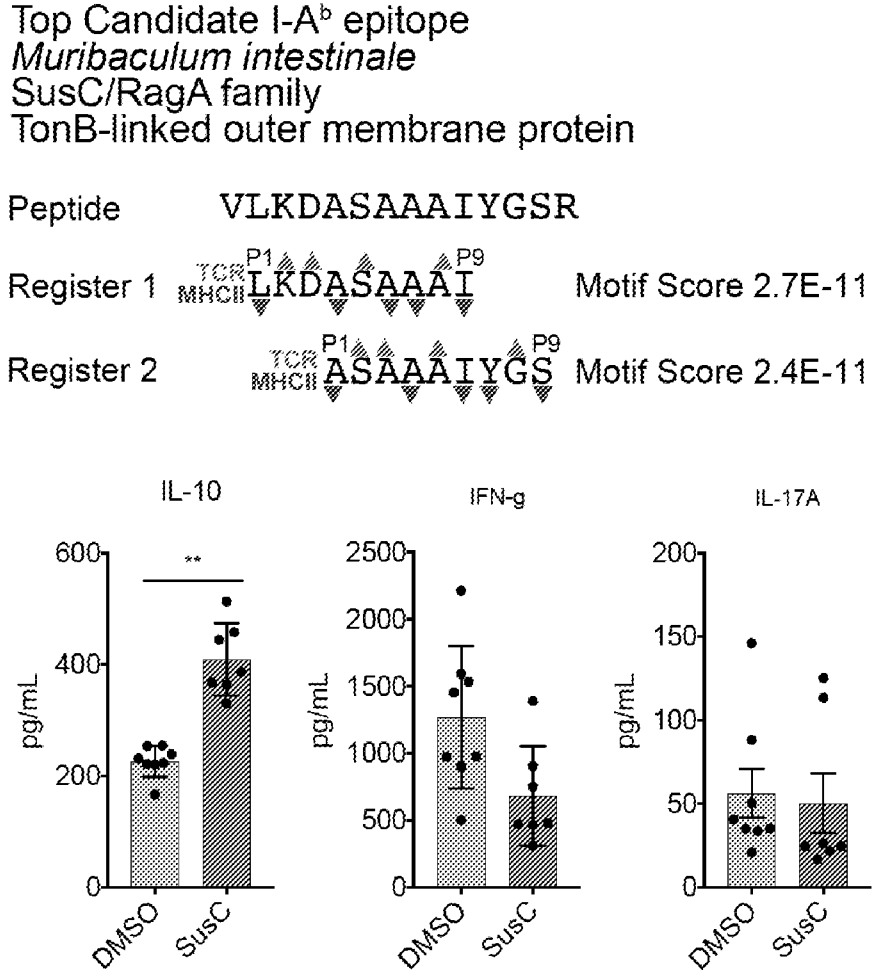



FIG. 15B




PCT/US2018/039843

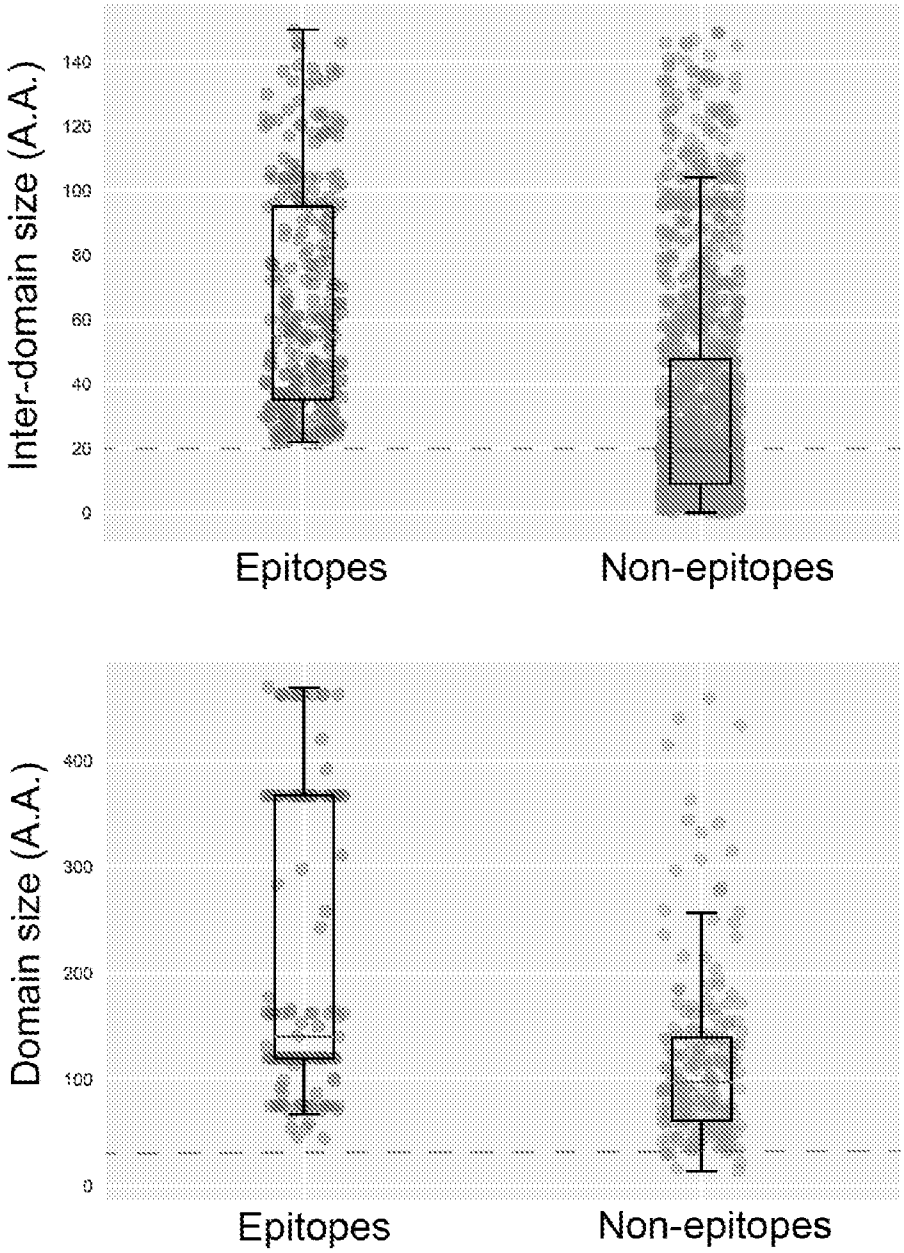

Inter-domain size (A.A.)
140
120
100
80
60
40
20
0
Epitopes
Non-epitopes
Domain size (A.A.)
400
300
200
100
0
Epitopes
Non-epitopes

FIG. 16

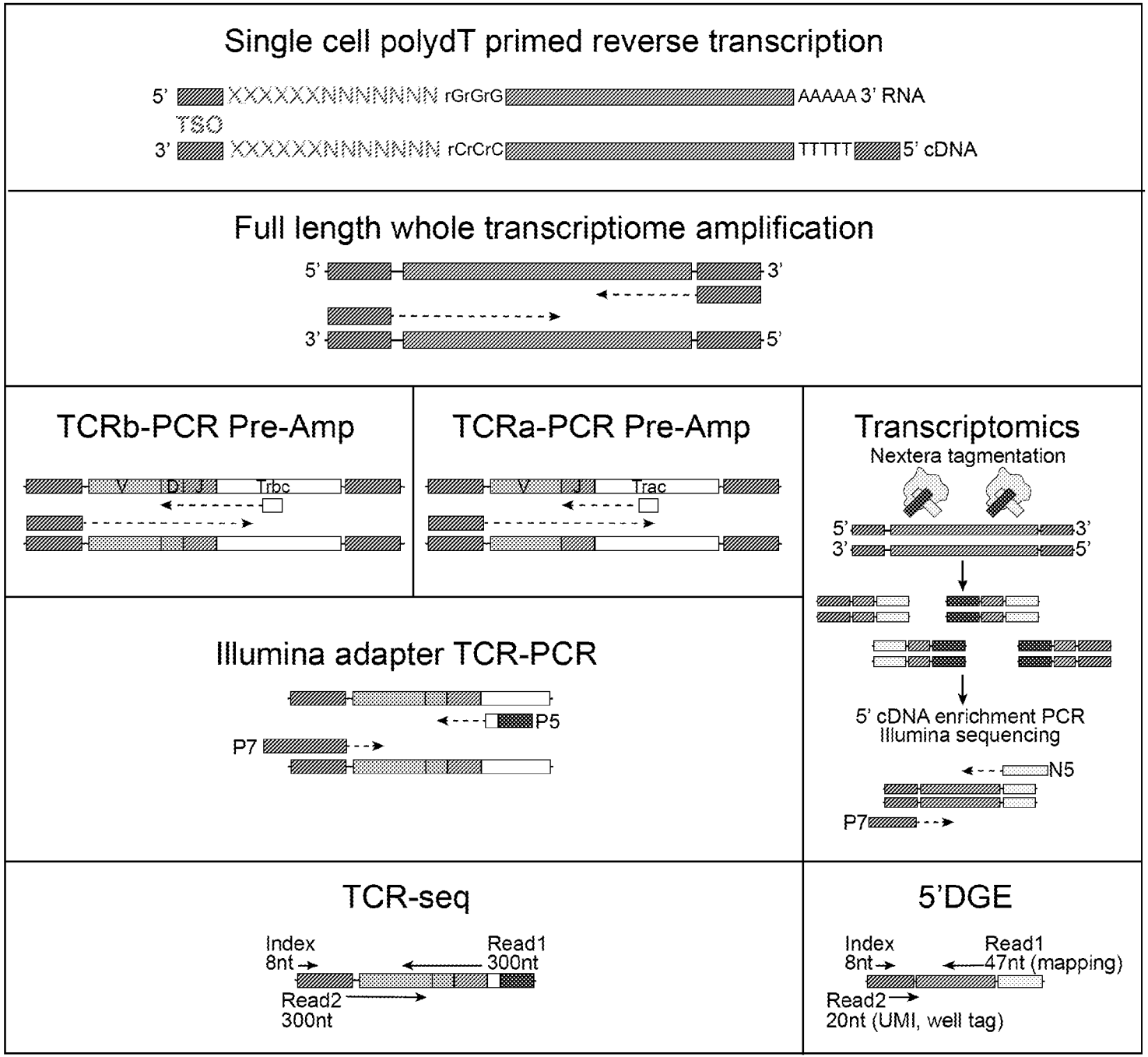



FIG. 17

WO 2019/006022

PCT/US2018/039843

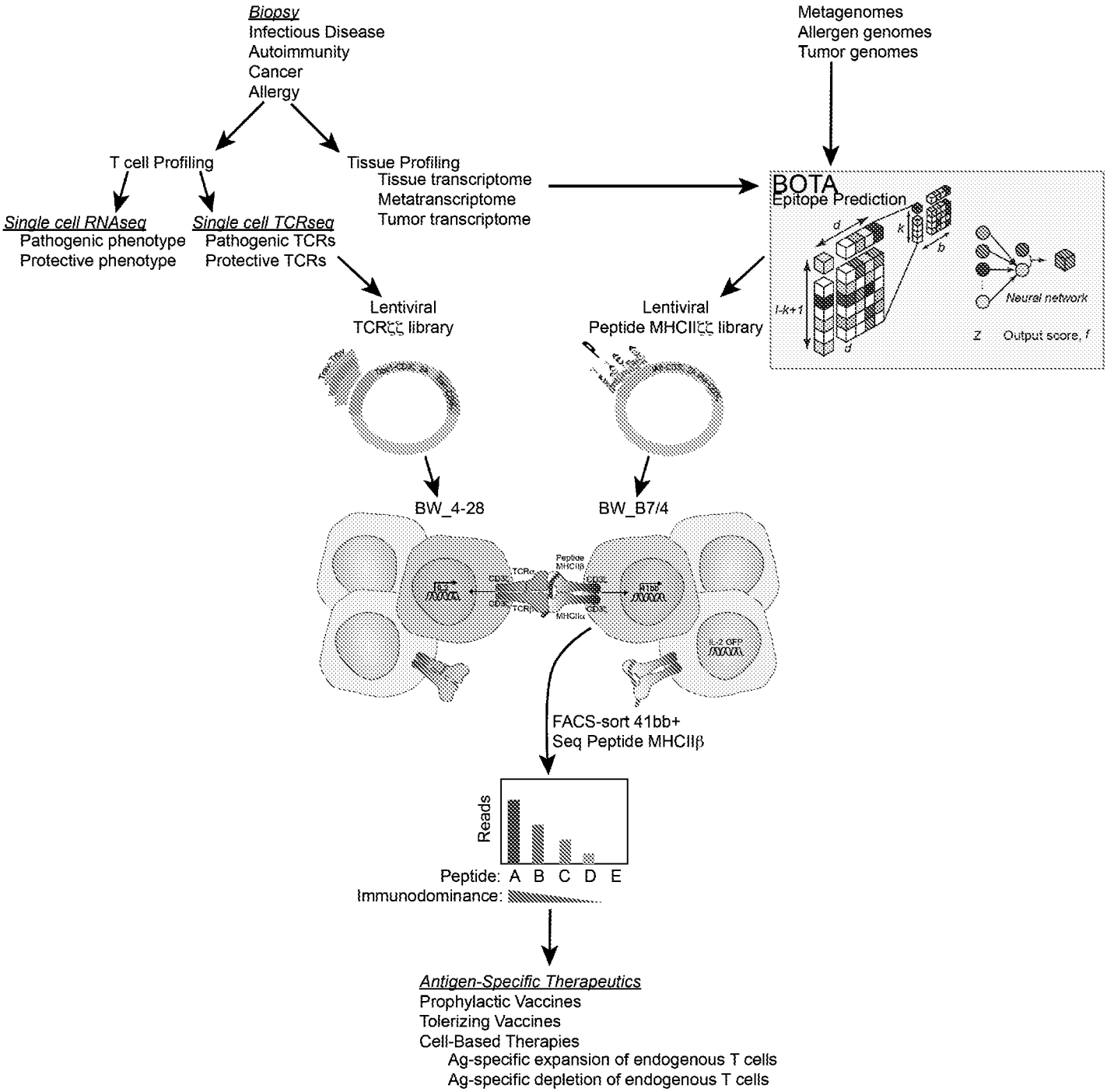

Biopsy
Infectious Disease
Autoimmunity
Cancer
Allergy
Metagenomes
Allergen genomes
Tumor genomes
T cell Profiling
Tissue Profiling
Tissue transcriptome
Metatranscriptome
Tumor transcriptome
BOTA
Epitope Prediction
Neural network
Output score, f
Single cell RNAseq
Pathogenic phenotype
Protective phenotype
Single cell TCRseq
Pathogenic TCRs
Protective TCRs
Lentiviral
TCRζζ library
Lentiviral
Peptide MHCIIζζ library
BW_4-28
BW_B7/4
FACS-sort 41bb+
Seq Peptide MHCIIβ
Reads
Peptide: A B C D E
Immunodominance:
Antigen-Specific Therapeutics
Prophylactic Vaccines
Tolerizing Vaccines
Cell-Based Therapies
Ag-specific expansion of endogenous T cells
Ag-specific depletion of endogenous T cells

FIG. 18

FIG. 19A

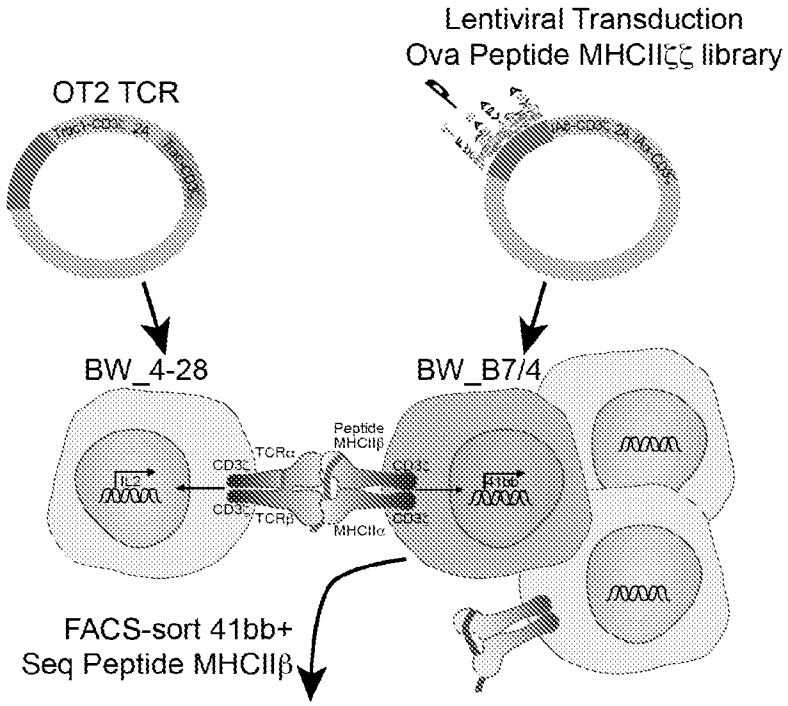



FIG. 19B

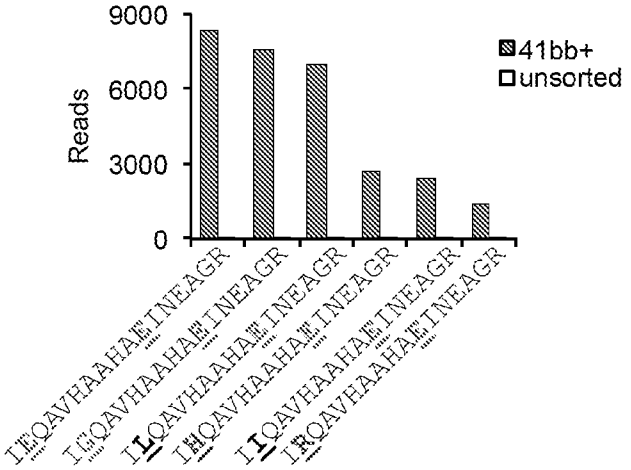
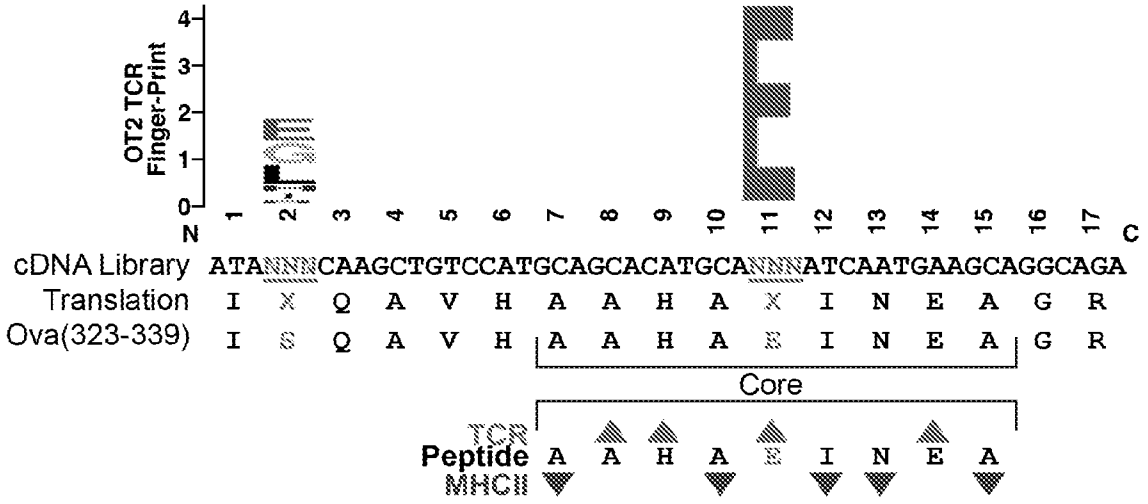



FIG. 19C



- **Probing the Host-Commensal Relationship to Reveal "Health Status" of the Immune System**

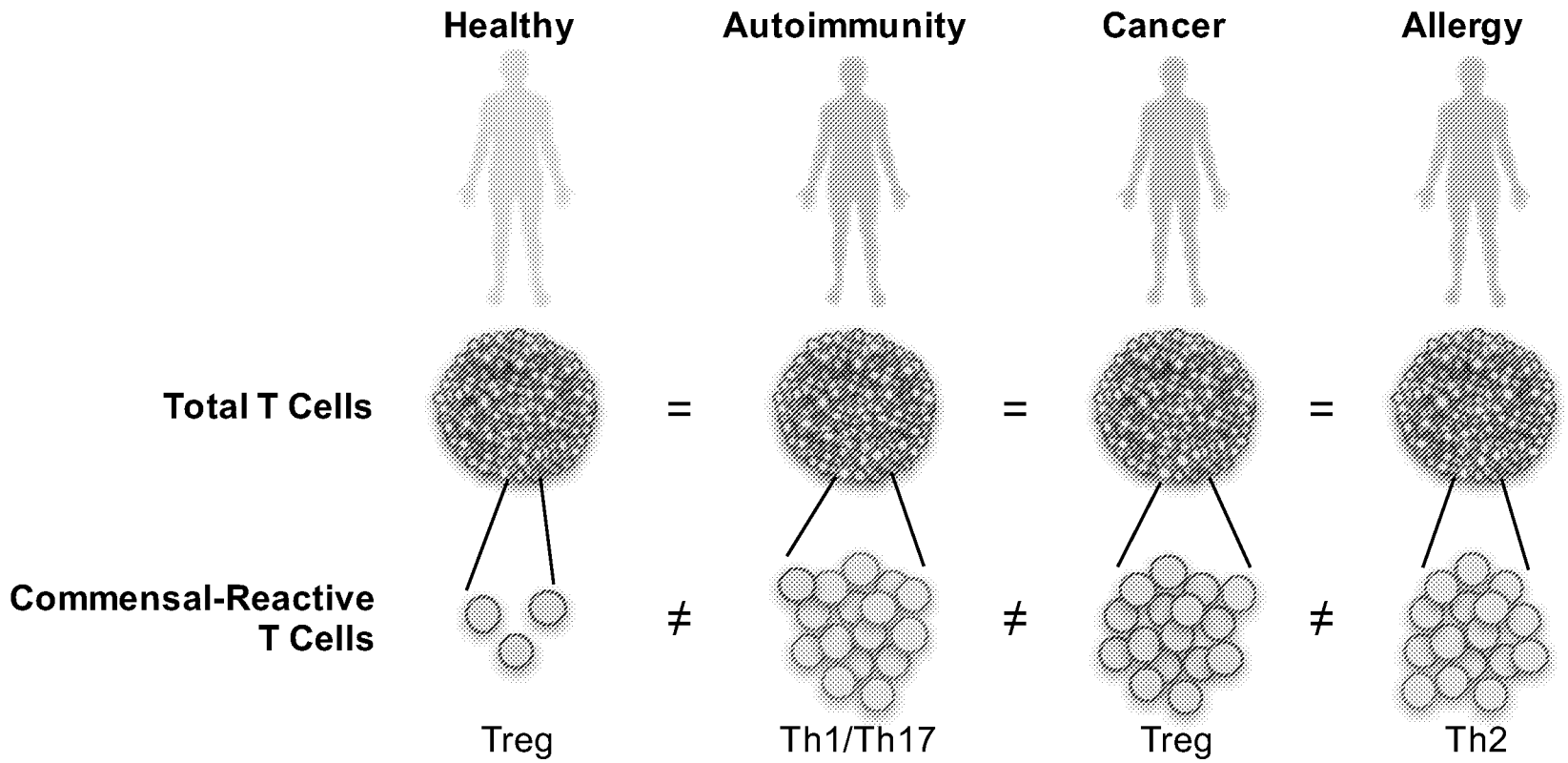



FIG. 20

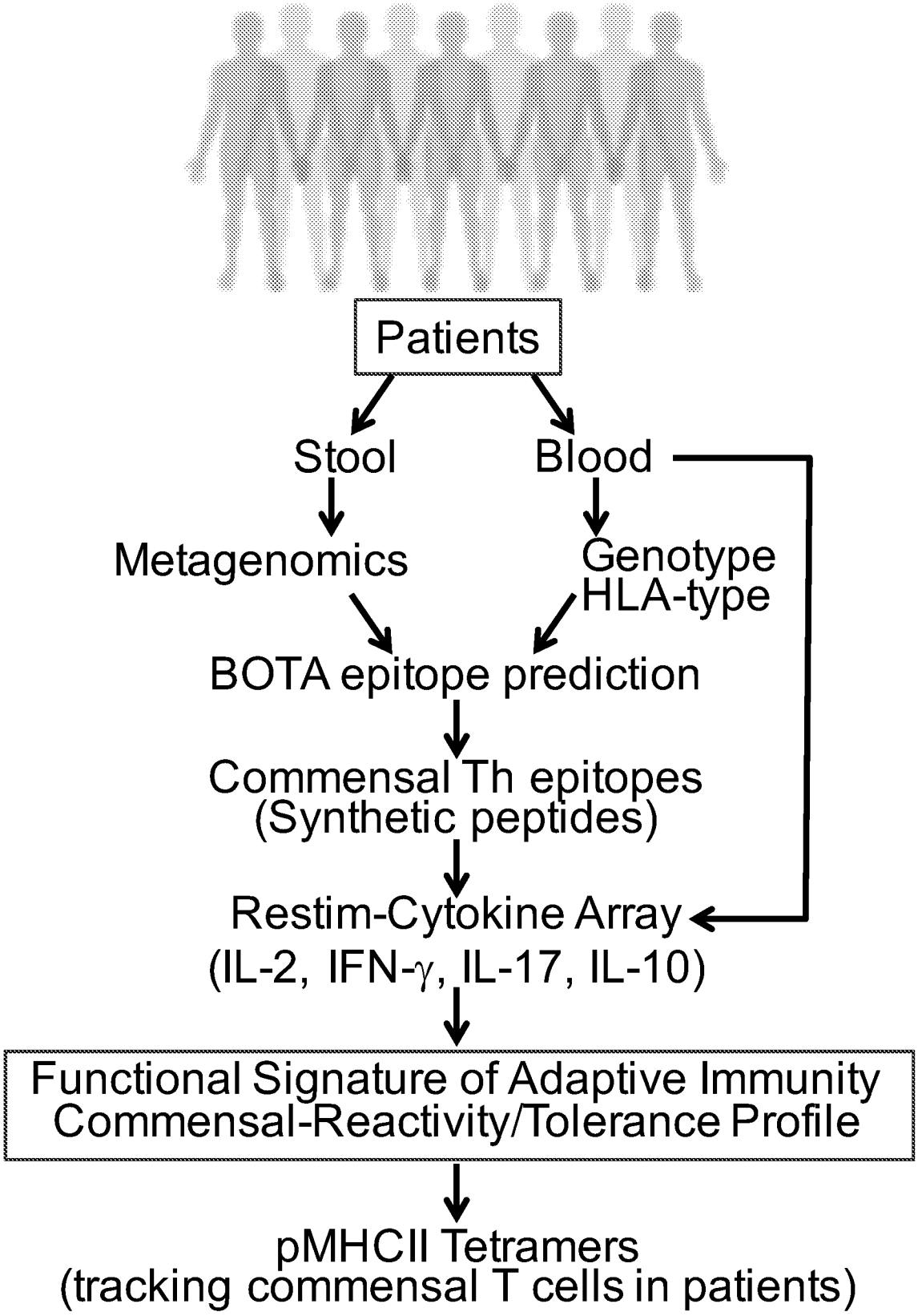

Patients
Stool
Blood
Metagenomics
Genotype
HLA-type
BOTA epitope prediction
Commensal Th epitopes
(Synthetic peptides)
Restim-Cytokine Array
(IL-2, IFN-γ, IL-17, IL-10)
Functional Signature of Adaptive Immunity
Commensal-Reactivity/Tolerance Profile
pMHCII Tetramers
(tracking commensal T cells in patients)

FIG. 21

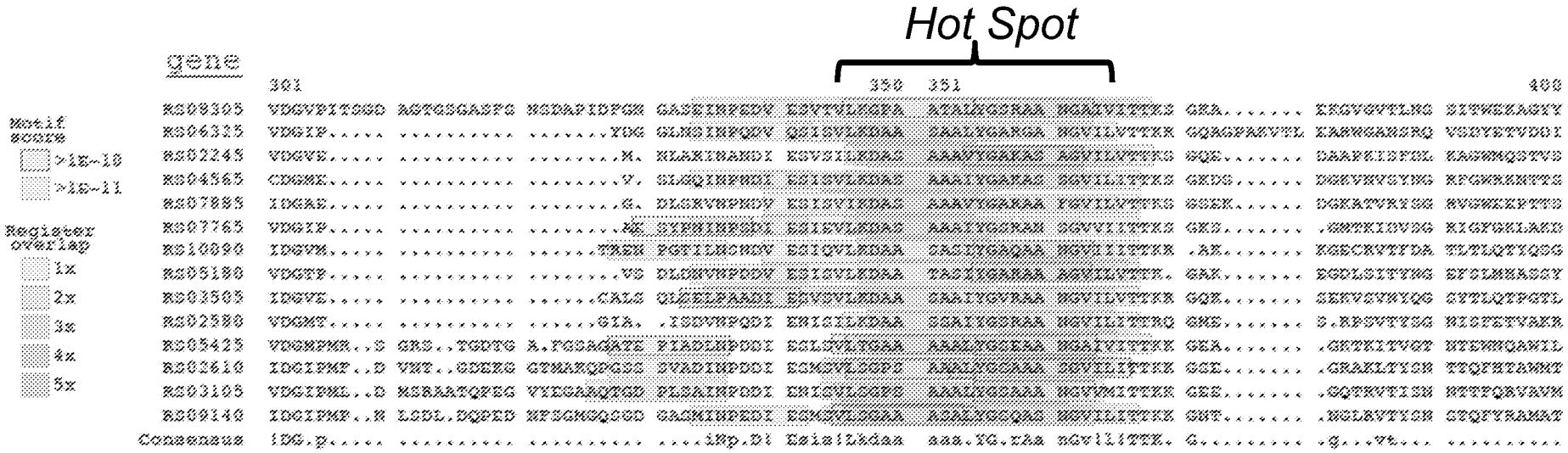
WO 2019/006022

PCT/US2018/039843

# Many SusC-like genes in the *bacteroidales* genome (including human-associated strains)

Muribaculum intestinale YL27
SusC/RagA TonB genes

Mouse I-A$^b$ (human orthologue HLA-DQ)
epitopes in conserved region




Tested Epitope: VLKDASAAAIYGSR
Associated with Treg reactivity in healthy mice

FIG. 22

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/39843

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 19/22; A61K 39/00, 39/02, 38/10, 35/15; C07K 7/08 (2018.01)
CPC - G06F 19/22; A61K 39/0011, 39/02, 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<br>---<br>A | WO 2017/106638 A1 (Gritstone Oncology, Inc.) 22 June 2017 (22 June 2017). Especially para [0082], [0083], [0086], [0087], [001 11], [001 15], [00120], [o0129], [00171], [00196], [0021 1], [00229], [02245], [00258], [00283], [00354]. | 1-8, 11, 12, 17-21, 23-25, 29-33, 36, 38, 44<br>---<br>9, 10, 13-16, 22, 26-28, 34, 37, 39, 40, 56-63<br>---<br>45-54 |
| Y | Jeannin et al. Outer membrane protein A (OmpA): a new pathogen-associated molecular pattern that interacts with antigen presenting cells-impact on vaccine strategies. Vaccine 19 December 2002 Vol 20 Suppl 4 Pages A23-A27. Especially abstract. | 9, 10 |
| Y | Zhu et al. Crystal structure of MHC class II I-Ab in complex with a human CLIP peptide: prediction of an I-Ab poptide-binding motif. J Mol Biol 28 Teb 2003 Vol 32G No 4 Pages 1157-1174. Especially abstract | 13-15 |
| Y | Androotta ot al. Aoouratc pan-spceific prediction of peptide-MI IC class II binding affinity with improved binding core identification. Immunogenetics 2015 November Vol 67 Pages 641-650; Especially [in NCBI Author's manuscript] pg 4 para 3-4, pg 5 para 2. | 16 |
| Y | Sun et al. Listeriolysin O as a strong immunogenic molecule for the development of new anti¬tumor vaccines. Hum Vaccin Immunother May 2013 Vol 9 No 5 Pages 1058-1068. Especially pg 1059 col 2 para 1, pg 1061 col 2 para 2, pg 1062 col 1 para 2. | 22, 27, 37, 40 |

☒ 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 alum;

"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

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 31 October 2018 | 06 DEC 2018 |

| Name and mailing address of the ISA/US | Authorized officer: |
|---|---|
| Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 2231 3-1450<br>Facsimile No. 571-273-8300 | Lee W. Young<br>PCT Helpdesk: 571-272-4300<br>PCT OSP: 571-272-7774 |

**INTERNATIONAL SEARCH REPORT**

International application No.
PCT/US 18/39843

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | US 5,468,481 A (Sharma et al.) 21 November 1995 (21.11.1995). Especially claims 1-3. | 26, 39 |
| Y | US 2015/0301044 A1 (Wisconsin Alumni Research Foundation) 22 October 2015 (22.10.2015). Especially para [0246], [0265], claim 1. | 28, 56-63 |
| Y | US 2012/0087862 A1 (Hood et al.) 12 April 2012 (12.04.2012). Especially SEQ ID NO: 47462 | 34 |
| A | BPS Bioscience. IL-2-Luciferase Reporter (Luc) - Jurkat Cell Line, 16 May 2017 [online]. [Retrieved on 31 October 2018]. Retrieved from the internet: <URL: https://web.archive.org/web/20170516091246/http://bpsbioscience.com/il2-jurkat-cell-line-60481 >. Especially pg 1. | 45-54 |
| A | Redmond et al. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Med 27 July 2016 Vol 8 No 80 Pages 1-12. Especially | 45-54 |

**INTERNATIONAL SEARCH REPORT**

International application No.
PCT/US 18/39843

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item l.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

a. ☒ forming part of the international application as filed:

☒ in the form of an Annex C/ST.25 text file.

☐ on paper or in the form of an image file.

b. ☐ furnished together with the international application under PCT Rule *12ter.* I(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

c. ☐ furnished subsequent to the international filing date for the purposes of international search only:

☐ in the form of an Annex C/ST.25 text file (Rule *13ter.* 1(a)).

☐ on paper or in the form of an image file (Rule 13/er. 1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:
GenCore ver 6.4.1 SEQ ID NO: 1

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/39843

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6\4(a).

Box No. Ill Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
— Go to Extra Sheet for continuation-----

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-54, 56-63, limited to peptide SEQ ID NO: 1 (Claims 1-34, 36-40, 44-54, 56-63)

**Remark on Protest**

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/39843

Continuation of Box III: Observations wherein Unity of Invention is lacking

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1 . In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I+: Claims 1-54, 56-63, drawn to a method of preparing one or more peptides for an immunological composition.

The method of preparing one or more peptides for an immunological composition will be searched to the extent that the peptide is the first named sequence, SEQ ID NO: 1 (claim 34). It is believed that claims 1-34, 36-40, 44-54, 56-63 read on this first named invention and thus these claims will be searched without fee to the extent that they encompass SEQ ID NO: 1. [note: Claims 35 and 41-43 are excluded from the first invention because claims 35 and claim 40 table 1 (instant application pg 63) EACH comprise SEQ ID NOs: 14980 -15027 only. Furthermore, claim 34, included as part of the first invention comprises some Listeria peptides among SEQ ID NOs: 1-14979, but not those listed in Table 1]. Additional peptides will be searched upon payment of additional fees. Applicant must specify the claims that encompass any additional elected peptides. Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be peptide SEQ ID NO: 14981 (claims 1-33, 35-54, 56-63).

Group II: Claims 55, 64-68, drawn to a method of constructing a deep neural network for identifying MHCII epitopes.

Group III+: Claims 69-74, drawn to a set of peptides comprising one or more peptides.

Group III+ will be searched upon payment of additional fee(s). The composition may be searched, for example, to the extent that the peptide encompasses SEQ ID NO: 14981 for an additional fee and election as such. It is believed that claims 70-74 read on this exemplary invention. Additional peptides will be searched upon the payment of additional fees. Applicants must indicate, if applicable, which claims read on this named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the '+' group(s) will result in only the first named invention to be searched/examined. An exemplary election would be the peptide comprises SEQ ID NO: 1 (claim 69).

The inventions listed as Groups I+, II, III+ do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Technical Features:

Group I+ has the special technical feature of a method of generating, using one or more processors, a set of candidate antigens from one or more input genome sequences, not required by Group II or III+.

Group I+ has the special technical feature of a method of generating, using one or more processors, a ranked set of antigenic epitopes using a deep neural network, not required by Groups II or III+.

Group II has the special technical feature of isolating MHCII-peptide complexes from antigen presenting cells (dendritic cells), not required by Groups I+ or III+

Group II has the special technical feature of isolating peptides from the MHCII-peptide complexes, not required by Groups I+ or III+.

Group III+ has the special technical feature of a composition of specifically defined peptide sequences, not required by Groups I+ or II.

No technical features are shared between the peptide sequences of Groups I+ and III+ and, accordingly, these groups lack unity a priori.

Additionally, even if Groups I+, II and III+ were considered to share the technical features of:
1. Group I+ inventions have the shared technical feature of claim 1.
2. Groups I+, II, III+ have the shared technical feature of peptides.
3. Group III+ inventions have the shared technical feature of claims 69-71 .

these shared technical features are previously disclosed by WO 2017/106638 A1 to Gritstone Oncology, Inc. (hereinafter "Gritstone") [published 22 June 2017], in view of the publication titled "Listeriolysin O as a strong immunogenic molecule for the development of new anti-tumor vaccines" by Sun et al. (hereinafter "Sun")[published in Hum Vaccin Immunother May 2013 Vol 9 No 5 Pages 1058-1068]

— continued on next sheet—

---continued from previous sheet------

As to shared technical features #1 and #2, Gritstone teaches (claim 1) a method of preparing one or more peptides for an immunological composition comprising:
a) generating, using one or more processors, a set of candidate antigens from one or more input genome sequences (para [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");
b) generating, using the one or more processors, a ranked set of antigenic epitopes from the candidate antigens using a deep neural network (para [0087]; "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"),
and
c) formulating an immunological composition comprising one or more of the identified epitopes (para [00120]; " 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").

As to claims 69 and 70, Gritstone teaches a peptide set for training a MHCII neural network (para [0098]; "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"; para [0099]; "The samples can also include cell lines engineered to express a single MHC class I or class II allele").

As to claim 71, Sun teaches a Listeria vaccine comprising one or more peptides (pg 1061 col 2 para 2; To date, three immunodominant epitopes have been determined by different experiments. As shown in Figure 1B, these include one dominant cytotoxic T lymphocyte (CTL) epitope, LL091-99 (residues 91-99), and two typical CD4+ T cell epitopes, LLO189-201 (residues 189-201), and LL0215-226 (residues 215-226)"; pg 1060 fig 1; sequences indicated).

As the shared technical features were known in the art at the time of the invention, they cannot be considered common special technical features that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Groups I+, II, III+ lack unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature.