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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 November 2018 (15.11.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/209223 A1

(51) International Patent Classification:
C07K 9/00 (2006.01) *C12Q 1/06* (2006.01)
C07K 16/00 (2006.01) *G01N 33/68* (2006.01)
C12Q 1/68 (2018.01)

(21) International Application Number:
PCT/US2018/032304

(22) International Filing Date:
11 May 2018 (11.05.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/505,537 12 May 2017 (12.05.2017) US

(71) Applicant: **THE BOARD OF REGENTS OF THE UNIVERSITY OF TEXAS SYSTEM** [US/US]; 210 West 7th Street, Austin, TX 78701 (US).

(72) Inventors: **OSTMEYER, Jared**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **COWELL, Lindsay**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **CHRISTLEY, Scott**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **ROUNDS, William**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **TOBY, Inmary**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **MONSON, Nancy**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US). **GREENBERG, Benjamin**; c/o UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390 (US).

(74) Agent: **HIGHLANDER, Steven, L.**; Parker Highlander PLLC, 1120 S. Capital of Texas Highway, Bldg. One, Suite 200, Austin, TX 78746 (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).

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))*

(54) Title: BIOCHEMICAL MOTIF IN CDR3 OF ANTIBODY SEQUENCES DIAGNOSIS AND PATIENTS WITH RELAPSING-REMITTING MULTIPLE SCLEROSIS

(57) Abstract: The present disclosure relates to the identification and application of adaptive immune receptor sequences that indicate an elevated risk of developing an adaptive immune receptor-related disease, such as an autoimmune disease (*e.g.*, multiple sclerosis). The biomarkers and methods can further aid in the prophylaxis and treatment of such diseases.

WO 2018/209223 A1

BIOCHEMICAL MOTIF IN CDR3 OF ANTIBODY SEQUENCES DIAGNOSIS AND PATIENTS WITH RELAPSING-REMITTING MULTIPLE SCLEROSIS

DESCRIPTION

PRIORITY CLAIM

This application claims benefit of priority to U.S. Provisional Application Serial No. 62/505,537, filed May 12, 2017, the entire contents of which are hereby incorporated by reference.

BACKGROUND

I. Field

The present disclosure relates to the fields of machine learning, immunology, medicine, and molecular biology. More particularly, it addresses a method for identifying disease biomarkers by examining the sequence of adaptive immune receptor complementary determining region (CDR) segments at the sequence level.

II. Related Art

Lymphocytes express immune receptors on their cell surface, the genes of which are somatically generated in developing lymphocytes through a DNA recombination process known as V(D)J recombination. V(D)J recombination assembles variable (V), diversity (D), and joining (J) gene segments into mature, composite genes. The diversity of gene sequences generated by V(D)J recombination is huge as a result of varying combinations of V, D, and J gene segments, as well as sequence modifications (*e.g.*, exonucleolytic activity and non-templated nucleotide addition) at the junctions of rearranged gene segments. As a result, each individual has millions of unique immune receptor genes. Somatic generation of a tremendously diverse repertoire of immune receptors enables effective immune responses against an essentially infinite array of antigens, such as those derived from pathogens or tumors, but it can also lead to detrimental effects, such as autoimmune responses and organ rejection following transplantation. The composition of immune repertoires shifts in response to such immunological events, and thus reflects previous and ongoing immune responses.

Deep sequencing of immune repertoires has made it possible to comprehensively profile the clonal composition of lymphocyte populations, opening the door for novel approaches to diagnose and prognosticate diseases with a driving immune component by identifying repertoire sequence patterns associated with important clinical phenotypes.

Recent studies support the feasibility of this approach. Patterns in the relative abundances of V gene segment types in a repertoire have been observed in association with various autoimmune diseases [1-3], as well as with metastasis-free/progression-free survival in basal-like and HER2-enriched breast cancer subtypes and the immunoreactive ovarian cancer subtype [4]. Repertoire diversity has been associated with prognosis in gastric cancer [5] and with outcome following Ipilimumab treatment for metastatic melanoma [6]. The inventors have demonstrated that VH4-containing genes in B cell repertoires from the cerebrospinal fluid of RRMS patients have higher replacement mutation frequencies at six codons than those in healthy controls [2, 7]. The sum of Z scores across the six codons can distinguish RRMS patients from those with other neurological diseases (OND) [7].

The methods applied to date for associating repertoire patterns with clinical phenotypes have focused on repertoire-level features, ignoring the vast amounts of information available in the millions of individual immune receptors comprising a repertoire. This has been due to difficulties accounting for the tremendous diversity of immune repertoires and the lack of methods for mapping the large number of individual sequences in a repertoire to a single phenotype label. The inventors have developed a novel method that addresses both limitations by combining widely used machine learning methods with innovative approaches for accommodating the extraordinary sequence diversity of immune receptors and for aggregating the set of predictions made for each sequence in a repertoire.

SUMMARY

Thus, in accordance with the present disclosure, there is provided a method of identifying a disease biomarker from adaptive immune receptor sequences comprising (a) obtaining the sequence encoding one or more adaptive immune receptors from a plurality of immune cells obtained from (i) a plurality of subjects having a given disease and (ii) a plurality of control subjects; (b) assessing the following biochemical properties for each amino acid lying in a plurality of arbitrarily defined regions or subregions within said one or more adaptive immune receptor sequences:

(i) polarity;
(ii) secondary structure;
(iii) molecular volume;
(iv) codon diversity, and
(v) electrostatic charge,

(c) selecting one or more regions or subregions within said one or more adaptive immune receptor sequences; (d) scoring each region or subregion based on the biochemical properties using a parameterized detector function; (e) aggregating the scores from a patient's plurality of said regions or subregions to predict a patient diagnosis; and (f) adjusting the parameters of the scoring function to yield the correct diagnosis for each patient in the example data, thereby identifying an adaptive immune receptor-related disease biomarker. The method may further comprise assessing the regions or subregions in step (b) in combination with the logarithm of the relative abundance (also known as frequency count), where the relative abundance is either:

(i) the relative abundance of the most abundant receptor containing the subregions or regions, or
(ii) the relative abundance of each receptor containing the subregions or regions, or
(iii) the relative abundance of the subregions or regions, which can be calculated by summing the abundances of all the CDR3 sequences containing the subregions or regions and dividing by the total count of all subregions or regions.

Step (a) may comprise amplification of said sequence and/or any high-throughput sequencing platform, including but not limited to 454 or Illumina sequencers. The disease may be a human or animal disease, syndrome, or disorder in which lymphocytes play a role,

such as multiple sclerosis. The immune coding region may be a full length antibody light or heavy chain, an antibody light or heavy chain variable region, or one, two, three, four, five or six CDRs. Only heavy chain CDR3 may be analyzed. CDR coding sequence or sequences may be obtained from a VH4 immunoglobulin and/or, the one or more CDRs may be light chain CDRs. The one or more immune sequences may be a T cell receptor sequence, such as a TCR alpha, a TCR beta, a TCR gamma, or a TCR delta chain. The one or more subregions each consist of between 5 and 10 codons. The one or more subregions each consist of 6 codons. The detector function may be a logistic regression function, or other than a logistic regression function The scores from a patient's plurality of said regions or subregions may be aggregated together, such as by taking the highest score among the plurality of scores, or by aggregating the scores using a generalized f-mean (also called a Kolmogorov mean) where the function is an exponential function. The biomarker may be used to diagnose and/or treat a patient.

In another embodiment, there is provided a method of identifying a subject as having or at risk of developing multiple sclerosis comprising (a) obtaining the sequence encoding one or more heavy chain CDR3s from a plurality of B cells obtained from a subject; (b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

DFNWFD (SEQ ID NO: 1)
IMKWFD (SEQ ID NO: 2)
DGSWAE (SEQ ID NO: 3)
DVWKAP (SEQ ID NO: 4)
DFWNEV (SEQ ID NO: 5)
RQRYLD (SEQ ID NO: 6)
DKNWLD (SEQ ID NO: 7)
NCHPFD (SEQ ID NO: 8)
HLNWFD (SEQ ID NO: 9)
QLFWFD (SEQ ID NO: 10)
EPQDAF (SEQ ID NO: 11)
LYHYDS (SEQ ID NO: 12)
DYWYLD (SEQ ID NO: 13)
DYWYFD (SEQ ID NO: 14)
WYLDLW (SEQ ID NO: 15)
WYFDLW (SEQ ID NO: 16)

EEQWLA (SEQ ID NO: 17)
KQQQRF (SEQ ID NO: 18)
DYSYFD (SEQ ID NO: 19)
SEWYID (SEQ ID NO: 20)
QTQSIV (SEQ ID NO: 21)
DCHYFD (SEQ ID NO: 22)
DWEWLL (SEQ ID NO: 23)
DVEWLL (SEQ ID NO: 24)
WEWLLF (SEQ ID NO: 25)
EWLFFD (SEQ ID NO: 26)
EWLLFD (SEQ ID NO: 27)
DLHHHY (SEQ ID NO: 28)
DLHCHY (SEQ ID NO: 29)
HYHYVM (SEQ ID NO: 30)
DLHYHY (SEQ ID NO: 31)
ELHYHY (SEQ ID NO: 32)
HHHYGM (SEQ ID NO: 33)
HPHDAF (SEQ ID NO: 34)
FCHPHD (SEQ ID NO: 35)
DAFDLW (SEQ ID NO: 36)
KFWDLL (SEQ ID NO: 37)
AIRHSD (SEQ ID NO: 38)
AVRHSD (SEQ ID NO: 39)
HLLLLH (SEQ ID NO: 40)
REHMAV (SEQ ID NO: 41)
WYLDLW (SEQ ID NO: 42)
WYFDLW (SEQ ID NO: 43)
EYFQHW (SEQ ID NO: 44)
HTNFDD (SEQ ID NO: 45)
WYFYLW (SEQ ID NO: 46)
HWRHCS (SEQ ID NO: 47)
HVRHCS (SEQ ID NO: 48)
SFHFDS (SEQ ID NO: 49)
ARHWRH (SEQ ID NO: 50)

HGRHCS (SEQ ID NO: 51)

HYYMDV (SEQ ID NO: 52)

and (c) identifying said subject as having or at risk of developing multiple sclerosis when one of more of said sequences is identified. The CDR coding sequences may be obtained from a VH4 immunoglobulin. Step (a) may comprise amplification of said sequence, and/or any high-throughput sequencing platform including but not limited to 454 or Illumina sequencers. The method may further comprise providing to said subject a therapeutic or prophylactic treatment for multiple sclerosis.

In yet another embodiment, there is provided a method of identifying a subject as having or at risk of developing colorectal cancer comprising (a) obtaining the sequence encoding one or more beta chain CDR3s from a plurality of T cells obtained from a subject; (b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

MGRM (SEQ ID NO: 53)

IRQM (SEQ ID NO: 54)

ENRI (SEQ ID NO: 55)

GRHM (SEQ ID NO: 56)

IRDM (SEQ ID NO: 57)

RGKM (SEQ ID NO: 58)

IGRM (SEQ ID NO: 59)

INKI (SEQ ID NO: 60)

HREF (SEQ ID NO: 61)

RRTM (SEQ ID NO: 62)

ERRM (SEQ ID NO: 63)

ERRM (SEQ ID NO: 64)

HNRM (SEQ ID NO: 65)

IRKE (SEQ ID NO: 66)

HGRM (SEQ ID NO: 67)

YREF (SEQ ID NO: 68)

WKDY (SEQ ID NO: 69)

MYRE (SEQ ID NO: 70)

YREV (SEQ ID NO: 71)

ERFY (SEQ ID NO: 72)

RERF (SEQ ID NO: 73)

MRGM (SEQ ID NO: 74)

ERSI (SEQ ID NO: 75)

IRQF (SEQ ID NO: 76)

RRHI (SEQ ID NO: 77); and

(c) identifying said subject as having or at risk of developing colorectal cancer when one of more of said sequences is identified.

Step (a) may comprise amplification of said sequence, and/or any high-throughput sequencing platform including but not limited to 454 or Illumina sequencers. The method may further comprise providing to said subject a therapeutic or prophylactic treatment for cancer. The subject may be suspected of having cancer, such as colorectal cancer. The subject may have previously been diagnosed as having cancer, such as colorectal cancer. The method may further comprise performing steps (a)-(c) a second time after said treatment to assess a change in the T cell repertoire.

In still a further embodiment, there is provided a method of identifying a subject as having or at risk of developing breast cancer comprising (a) obtaining the sequence encoding one or more beta chain CDR3s from a plurality of T cells obtained from a subject; (b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

LSRG (SEQ ID NO: 78)

LSRS (SEQ ID NO: 79)

RSNQ (SEQ ID NO: 80)

LSYE (SEQ ID NO: 81)

ASYN (SEQ ID NO: 82)

AGNQ (SEQ ID NO: 83)

GSYN (SEQ ID NO: 84)

ASNQ (SEQ ID NO: 85)

LCNN (SEQ ID NO: 86)

ASYE (SEQ ID NO: 87)

SSYN (SEQ ID NO: 88)

LPRD (SEQ ID NO: 89)

SSYN (SEQ ID NO: 90)

LDGQ (SEQ ID NO: 91)

PSNQ (SEQ ID NO: 92)

ASNE (SEQ ID NO: 93)

AYNQ (SEQ ID NO: 94)
AAYN (SEQ ID NO: 95)
SSPH (SEQ ID NO: 96)
DSNQ (SEQ ID NO: 97)
SSNN (SEQ ID NO: 98)
SSYE (SEQ ID NO: 99)
ASNQ (SEQ ID NO: 100)
SSYN (SEQ ID NO: 101)
ASRD (SEQ ID NO: 102)
SSKD (SEQ ID NO: 103); and

(c) identifying said subject as having or at risk of developing breast cancer when one of more of said sequences is identified.

Step (a) may comprise amplification of said sequence, and/or any high-throughput sequencing platform including but not limited to 454 or Illumina sequencers. The method may further comprise providing to said subject a therapeutic or prophylactic treatment for cancer. The subject may be suspected of having cancer, such as breast cancer. The subject may have previously been diagnosed as having cancer, such as breast cancer. The method may further comprise performing steps (a)-(c) a second time after said treatment to assess a change in the T cell repertoire.

As used herein the specification, "a" or "an" may mean one or more. As used herein in the claim(s), when used in conjunction with the word "comprising", the words "a" or "an" may mean one or more than one.

The use of the term "or" in the claims is used to mean "and/or" unless explicitly indicated to refer to alternatives only or the alternatives are mutually exclusive, although the disclosure supports a definition that refers to only alternatives and "and/or." As used herein "another" may mean at least a second or more.

Throughout this application, the term "about" is used to indicate that a value includes the inherent variation of error for the device, for the method being employed to determine the value, or that exists among the study subjects. Such an inherent variation may be a variation of ±10% of the stated value.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications

within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

5

BRIEF DESCRIPTION OF THE DRAWINGS

The following drawings form part of the present specification and are included to further demonstrate certain aspects of the present disclosure. The disclosure may be better understood by reference to one or more of these drawings in combination with the detailed description of specific embodiments presented herein.

FIGS. 1A-D. (FIG. 1A) Study Overview (FIG. 1A) B cells are collected from patient cerebrospinal fluid. (FIG. 1B) DNA is extracted and next generation sequencing is used to sequence immunoglobulin heavy chain loci expressing IGHV4 rearrangements. (FIG. 1C) Snippets of amino acid sequence taken from the CDR3 are converted into a set of chemical features using Atchley factors. (FIG. 1D) The chemical features are scored by a detector function. The detector function used in this study is the same function used in logistic regression. A positive diagnosis (for RRMS) is flagged whenever a high scoring snippet is found. Values for the weights on each Atchley factor as well as the bias term are determined by maximizing the likelihood of obtaining the correct diagnoses on a training set of patients.

FIGS. 2A-B. Workflow for Model Selection and Parameter Fitting (FIG. 2A) The diagram shows how training data is used to train and evaluate multiple hypotheses. The model that gives the best classification accuracy on the exhaustive 1-holdout cross-validation constitutes the lead hypothesis. (FIG. 2B) The diagram shows how data is used to train and test the lead hypothesis. The best performing model is refitted to all the samples in the training data, and then used to score samples from the validation data set.

FIGS. 3A-D. Classification Accuracy and Receiver Operating Characteristic (ROC) Curves (FIG. 3A) Classification accuracy for the best performing model obtained via exhaustive 1-holdout cross-validation on training data. 87% of patients were correctly classified. (FIG. 2B) The corresponding ROC curve shows true and false positive rates for different thresholds of a positive diagnosis based on the highest snippet score. The area under the curve is 0.86. (FIG. 2C) Classification accuracy of the best performing model on the validation data. 72% of patients were correctly classified. (FIG. 2D) The corresponding ROC curve is shown. The area under the curve is 0.75.

FIG. 4. Illustration of the Classifier Weights. For each of the five Atchley factors, the weights for the model fit on all 23 training samples are shown for the six residue

positions. Positive weight values are shown in red pointing up, and negative weight values are shown in blue pointing down. The length of the arrow corresponds to the weights magnitude.

FIG. 5. The Highest Scoring Snippet from each Patient in the Training Data Set. The snippets were scored with the model trained on all 23 subjects. (Left) Location of the snippet is shown in its CDR3 sequence using yellow highlight. (Right) The Atchley factor values are shown for each snippet in the five boxes. Each box corresponds to one Atchley factor. The columns in each box correspond to the snippet positions. (SEQ ID NOS: 104-126)

FIG. 6. Histograms of Snippet Scores for all Snippets in the Training Data Set. The snippets were scored with the model trained on all 23 subjects. The lighter bars indicate the distribution of snippet scores from RRMS patients. The dark bars indicate the distribution of snippet scores from OND patients. Only a few snippets score above 0.5, which is diagnostic of RRMS.

FIG. 7. An example of how to calculate the number of ways a snippet can be encoded. (SEQ ID NO: 47)

FIGS. 8A-D. (FIG. 8A) X-ray crystallographic structure of a human T-cell receptor β-chain (TCRB) bound to an antigen (ANTIGEN). The portion of the CDR3 (CDR3) in direct contact (≥ 5Å) with the antigen lies in the middle of the CDR3 and forms a contiguous strip (DIRECT CONTACT). The MHC complex and α-chain are omitted for clarity. (FIG. 8B) CDR3 sequences extracted from 57 X-ray crystallographic structures of human T-cell receptors bound to an antigen. Residues in direct contact with the antigen (shaded) are used to align the sequences. The first and last 3 residues of each CDR3 almost never contact the antigen. Contact residues tend to form a contiguous strip of about four residues, as indicated in the bar chart. (SEQ ID NOS: 127-183) (FIG. 8C) To profile the specificity of a CDR3 sequence, the CDR3 is cut into every possible snippet of 4 residues excluding the first and last 3 residues. (SEQ ID NOS: 232-237) (FIG. 8D) Each snippet is converted into a biochemical representation. For each residue, there are 5 Atchley factor values describing the residues biochemical properties. Using a snippet of 4 residues leads to 20 biochemical values.

FIG. 9. Workflow for model selection and parameter fitting. Diagram shows how the data are used to train and validate each model. The performance of each model is assessed by a patient-holdout cross-validation, where the tumor and healthy tissue

from the same patient are excluded for validation. Data from the remaining N-1 patients is used to fit the model. The fitting procedure is run for 2,500 steps and restarted with different initial weights 250,000 times. The best fit to the training samples is used to evaluate the excluded validation data.

FIGS. 10A-C. Colorectal cancer results. (FIG. 10A) Classification accuracy obtained by a patient-holdout cross-validation, where the tumor and healthy tissue from the same patient are excluded for validation. 93% of the samples are correctly classified, and 100% of the tumor samples score above the patient matched control tissue. (FIG. 10B) Illustration of the classifier weights after fitting the model to all 14 patients. For each of the five Atchley factors, the weights are shown for the four residue positions. The weight for the log-frequency of the snippet is also shown. Positive weight values are shown in arrow pointing up, and negative weight values are shown in arrow pointing down. The length of the arrow corresponds to the weight's magnitude. (FIG. 10C) All snippets with a score above 0.5 (middle column, shaded) shown for each of the 14 patients (leftmost column). Each snippet is embedded in its respective CDR3. When the snippet appears in multiple CDR3 sequences, the CDR3 with the largest relative abundance is shown. The CDR3 sequences are ranked according to their relative abundance in the sample (rightmost column). (SEQ ID NOS: 184-206)

FIGS. 11A-C. Breast cancer results. (FIG. 11A) Classification accuracy obtained by a patient-holdout cross-validation, where the tumor and healthy tissue from the same patient are excluded for validation. 94% of the samples are correctly classified, and 100% of the tumor samples score above the patient matched control tissue. (FIG. 11B) Illustration of the classifier weights after fitting the model to all 16 patients. For each of the five Atchley factors, the weights are shown for the four residue positions. The weight for the log-frequency of the receptor is also shown. Positive weight values are shown in the arrow pointing up, and negative weight values are shown in the arrow pointing down. The length of the *arrow* corresponds to the weight's magnitude. (FIG. 11C) All snippets with a score above 0.5 (middle column, shaded) shown for each of the 16 patients (leftmost column). Each snippet is embedded in its respective CDR3. When the snippet appears in multiple CDR3 sequences, the CDR3 with the largest relative abundance is shown. The CDR3 sequences are ranked according to their relative abundance in the sample (rightmost column). (SEQ ID NOS: 207-231)

DETAILED DESCRIPTION

As discussed above, multiple sclerosis (MS) is an autoimmune disease of the central nervous system characterized by a loss of myelin coating the neural exons. Because the underlying antigens remain unknown, no test for antibody autoreactivity exists. To accurately diagnose MS without the use of a known antigen, methods should leverage the information stored in a patient's antibody repertoire. However, existing techniques for performing statistical classification, such as logistic regression, are only able to map a fixed number of features from a patient's antibody repertoire to their diagnosis. The inventors therefore developed a novel method that would enable classification of sets of sequences (repertoires) based on features of all the individual sequences in the repertoire.

The inventors applied their methods to RRMS, a subtype of multiple sclerosis (MS). MS is an autoimmune disease that is notoriously difficult to diagnose. It is believed to be the result of immune cells attacking the myelin insulation around axons, leaving patients with physical and cognitive impairments. Unfortunately, there are no symptoms, physical findings, or lab tests that provide a definitive MS diagnosis. Patients have to demonstrate findings consistent with MS and simultaneously have alternative diagnoses be excluded [8]. Thus, reaching an MS diagnosis can be a slow process, but early detection is needed, because prompt intervention can significantly slow the progression of the disease [9].

The inventors applied their methods to B cell receptor (BCR) heavy chain genes to develop a statistical classifier that assigns patients to one of two diagnosis categories, RRMS or OND, based on the BCR heavy chain biochemical features. The classifier has 87% accuracy by leave-one-out cross-validation on training data (N = 23) and 73% accuracy on unused data from a separate study (N = 102). These results demonstrate the utility of this new method for identifying repertoire-based signatures with diagnostic potential.

By taking advantage of next generation sequencing, antibody DNA from patients with relapsing remitting multiple sclerosis (RRMS) along with a group of control patients diagnosed with other neurological diseases (OND) can be sequenced and used to predict each patient's diagnosis from their antibody sequences, a new type of statistical classifier was developed which uses a detector function to flag a positive diagnosis based on the set of predictions from each sequence. Using the biochemical features encoded by the CDR3 of each antibody as input, the parameters of the detector function are fitted by maximizing the likelihood of correctly diagnosing each patient. Once the parameters of the detector function have been fitted, it can be used to diagnose new patients, and would do so in an accurate

fashion. The model developed here correctly classifies all 23 patients used as training data, 20/23 patients using a 1-holdout cross-validation analysis of the training data, and 73/102 patients from a separate study that serves as unused and unseen data. These and other aspects of the disclosure are set forth below.

I. Lymphocyte Isolation Procedures

Lymphocytes can be isolated as blood or cerebrospinal fluid, or from almost any tissue, such lymphoid organs including the thymus, bone marrow, lymph nodes, and mucosal-associated lymphoid tissues. Density centrifugation isolates a population of peripheral blood mononuclear cells (PBMCs) by separating the solution into layers of differing densities. PBMC layers contain mononuclear cells that have been depleted of red blood cells, leukocytes and granulocytes.

Biopanning isolates cell populations from solution. Cells of interest are bound to antibody-coated plastic surfaces, and unwanted cells are removed by treatment with specific antibody and complement.

Fluorescence-activated cell sorter (FACS) analysis detects and counts lymphocytes passing through a laser beam. The FACS is a flow cytometer that separates labelled cells based on differences in the light scattering and fluorescence of cells.

Upon isolation, lymphocytes may be characterized in terms of specificity, frequency and function. Frequently used assays include the ELISPOT, which measures the frequency of T cell response. It is similar to the ELISA assay in that antibodies bound to plastic wells are used to bind the cytokines secreted by T cells.

II. High Throughput Sequencing of Antibody Coding Regions

High-throughput (formerly "next-generation") sequencing applies to genome sequencing, genome resequencing, transcriptome profiling (RNA-Seq), DNA-protein interactions (ChIP-sequencing), epigenome characterization, and sequencing of PCR product. Resequencing is necessary, because the genome of a single individual of a species will not indicate all of the genome variations among other individuals of the same species.

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-

terminator methods. In ultra-high-throughput sequencing, 500,000 or more sequencing-by-synthesis operations may be run in parallel.

SMRT sequencing is based on the sequencing by synthesis approach. The DNA is synthesized in zero-mode wave-guides (ZMWs) – small well-like containers with the capturing tools located at the bottom of the well. The sequencing is performed with use of unmodified polymerase (attached to the ZMW bottom) and fluorescently labelled nucleotides flowing freely in the solution. The wells are constructed in a way that only the fluorescence occurring by the bottom of the well is detected. The fluorescent label is detached from the nucleotide upon its incorporation into the DNA strand, leaving an unmodified DNA strand. According to Pacific Biosciences (PacBio), the SMRT technology developer, this methodology allows detection of nucleotide modifications (such as cytosine methylation). This happens through the observation of polymerase kinetics. This approach allows reads of 20,000 nucleotides or more, with average read lengths of 5 kilobases. In 2015, Pacific Biosciences announced the launch of a new sequencing instrument called the Sequel System, with 1 million ZMWs compared to 150,000 ZMWs in the PacBio RS II instrument. SMRT sequencing is referred to as "third-generation" or "long-read" sequencing.

The DNA passing through the nanopore changes its ion current. This change is dependent on the shape, size and length of the DNA sequence. Each type of the nucleotide blocks the ion flow through the pore for a different period of time. The method does not require modified nucleotides and is performed in real time. Nanopore sequencing is referred to as "third-generation" or "long-read" sequencing, along with SMRT sequencing.

Early industrial research into this method was based on a technique called 'Exonuclease sequencing', where the readout of electrical signals occurring at nucleotides passing by alpha-hemolysin pores covalently bound with cyclodextrin. However the subsequently commercial method, 'strand sequencing' sequencing DNA bases in an intact strand.

Two main areas of nanopore sequencing in development are solid state nanopore sequencing, and protein based nanopore sequencing. Protein nanopore sequencing utilizes membrane protein complexes such as ∝-Hemolysin, MspA (Mycobacterium Smegmatis Porin A) or CssG, which show great promise given their ability to distinguish between individual and groups of nucleotides. In contrast, solid-state nanopore sequencing utilizes synthetic materials such as silicon nitride and aluminum oxide and it is preferred for its superior mechanical ability and thermal and chemical stability. The fabrication method is

essential for this type of sequencing given that the nanopore array can contain hundreds of pores with diameters smaller than eight nanometers.

The concept originated from the idea that single stranded DNA or RNA molecules can be electrophoretically driven in a strict linear sequence through a biological pore that can be less than eight nanometers, and can be detected given that the molecules release an ionic current while moving through the pore. The pore contains a detection region capable of recognizing different bases, with each base generating various time specific signals corresponding to the sequence of bases as they cross the pore which are then evaluated. Precise control over the DNA transport through the pore is crucial for success. Various enzymes such as exonucleases and polymerases have been used to moderate this process by positioning them near the pore's entrance.

III. Scoring of CDR Sequences

Every "snippet" from every CDR3 sequence in a patient's repertoire is scored by a detector function indicating if a snippet predicts RRMS. The inventors use a logistic function because of its widespread use and simplicity, and because it models the outcome of a two-category process. The first step is to compute a biased, weighted sum of the snippet's features, referred to as a logit.

$$\text{logit} = b_0 + W_1 \cdot f_1 + W_2 \cdot f_2 + \cdots + W_N \cdot f_N$$

For the DNA and amino acid sequence representations, the values f_1 through f_N represent the snippet residues. For the Atchley factor representation, the f_i represent the five Atchley factors from each residue in the snippet. For snippets of length six, $N = 30$. The bias term b_0 along with the weights W_i are the parameters of the model and are fit by maximum likelihood using gradient descent optimization techniques as described below. The same weights W_i and bias term b_0 are used for all snippets. Once the logit is computed, the value is passed through the sigmoid function to obtain a score between 0 and 1.

$$\text{score} = {}^{1}/_{1 + e^{-\text{logit}}}$$

A patient's snippet scores need to be aggregated into a single value to form a diagnosis. Because only a small fraction of BCRs in a patient's repertoire are expected to be disease related, it is necessary to capture a diagnosis even if only a few snippets have a high score. This is accomplished by assigning a positive diagnosis when even a single high scoring snippet is found (FIG. 1D). Assuming the output of the detector function represents a probability value between 0 and 1, the form of the model can be written as:

$$P(\text{ positive diagnosis} \mid \text{snip}_1, \text{snip}_2, \text{snip}_3, \ldots) = \text{Maximum}(\text{ score}_1, \text{score}_2, \text{score}_3, \ldots)$$

A probability > 0.5 indicates a positive diagnosis (RRMS), whereas a value < 0.5 indicates an OND diagnosis.

Specific values for the weights W_i and bias term b_0 in the detector function are determined using the patient diagnoses. The values must be chosen to maximize the likelihood that each predicted diagnosis is correct. To search for the optimal values, gradient optimization techniques are used. With these techniques, each parameter is iteratively adjusted along the gradient in a direction that maximizes the log-likelihood, which in turn maximizes the likelihood that each predicted diagnosis is correct. The initial value for the bias term b_0 is 0, and initial values for the weights are drawn at random according to $W_i \sim \mathcal{N}(0, N_{\text{features}}^{-1})$. Because the ADAM optimizer, a gradient descent based method, has been shown to work well on a wide range of optimization tasks, it is used here [11]. The ADAM optimizer is run for 2500 iterations with a step size of 0.01. The default values for the other ADAM optimizer settings are: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

A limitation of using a gradient descent based method is there is no guarantee of finding the globally optimal solution. Although the chosen detector function constitutes a linear model, the scores from every snippet are aggregated together in a non-linear fashion. Multiple local minima could exist. To address this, 10^5 runs of gradient descent, each starting from different initial parameters $W_i \sim \mathcal{N}(0, N_{\text{features}}^{-1})$, are used, and the best fit solution over all runs is used to diagnose new patients.

In addition, a recurrent neural network could be used as an alternative detector function. The advantage of this approach is that it can accommodate the variable length nature of the adaptive immune receptor sequences. The scores can also be aggregated into a single score using a generalized f-mean (also called a Kolmogorov mean) where the function is an exponential function.

IV. Immunity and Disease

A. Autoimmune Disease

An autoimmune disease is a condition arising from an abnormal immune response to a normal body part. There are at least 80 types of autoimmune diseases. Nearly any body part can be involved. Common symptoms include low grade fever and feeling tired. Often symptoms come and go.

The cause is generally unknown. Some autoimmune diseases such as lupus run in families, and certain cases may be triggered by infections or other environmental factors. Some common autoimmune disease include celiac disease, diabetes mellitus type 1, Sjogren's disease, Graves' disease, inflammatory bowel disease, multiple sclerosis, psoriasis, rheumatoid arthritis, and systemic lupus erythematosus. The diagnosis can be difficult to determine.

Treatment depends on the type and severity of the condition. Nonsteroidal anti-inflammatory drugs (NSAIDs) and immunosuppressants are often used. Intravenous Immunoglobulin may also occasionally be used. While treatment usually improves symptoms they do not typically cure the disease.

The human immune system typically produces both T-cells and B-cells that are capable of being reactive with self-antigens, but these self-reactive cells are usually either killed prior to becoming active within the immune system, placed into a state of anergy (silently removed from their role within the immune system due to over-activation), or removed from their role within the immune system by regulatory cells. When any one of these mechanisms fail, it is possible to have a reservoir of self-reactive cells that become functional within the immune system. The mechanisms of preventing self-reactive T-cells from being created takes place through Negative selection process within the thymus as the T-cell is developing into a mature immune cell.

Some infections, such as *Campylobacter jejuni*, have antigens that are similar (but not identical) to one's own self-molecules. In this case, a normal immune response to *C. jejuni* can result in the production of antibodies that also react to a lesser degree with peripheral nerve myelin (*e.g.*, Guillain-Barre Syndrome). A major understanding of the underlying pathophysiology of autoimmune diseases has been the application of genome wide association scans that have identified a degree of genetic sharing among the autoimmune diseases.

Autoimmunity, on the other hand, is the presence of self-reactive immune response (*e.g.*, auto-antibodies, self-reactive T-cells), with or without damage or pathology resulting from it. This may be restricted to certain organs (*e.g.*, in autoimmune thyroiditis) or involve a particular tissue in different places (*e.g.*, Goodpasture's disease which may affect the basement membrane in both the lung and the kidney).

B. Adaptive Immunity

The cells of the adaptive immune system are T and B lymphocytes; lymphocytes are a subset of leukocyte. B cells and T cells are the major types of lymphocytes. The human body has about 2 trillion lymphocytes, constituting 20–40% of white blood cells (WBCs); their total mass is about the same as the brain or liver. The peripheral blood contains 2% of circulating lymphocytes; the rest move within the tissues and lymphatic system.

B cells and T cells are derived from the same multipotent hematopoietic stem cells, and are morphologically indistinguishable from one another until after they are activated. B cells play a large role in the humoral immune response, whereas T cells are intimately involved in cell-mediated immune responses. In all vertebrates except Agnatha, B cells and T cells are produced by stem cells in the bone marrow.

T progenitors migrate from the bone marrow to the thymus where they are called thymocytes and where they develop into T cells. In humans, approximately 1–2% of the lymphocyte pool recirculates each hour to optimize the opportunities for antigen-specific lymphocytes to find their specific antigen within the secondary lymphoid tissues. In an adult animal, the peripheral lymphoid organs contain a mixture of B and T cells in at least three stages of differentiation:

naive B and naive T cells (cells that have not matured), left the bone marrow or thymus, have entered the lymphatic system, but have yet to encounter their cognate antigen,

effector cells that have been activated by their cognate antigen, and are actively involved in eliminating a pathogen.

memory cells – the survivors of past infections.

Adaptive immunity relies on the capacity of immune cells to distinguish between the body's own cells and unwanted invaders. The host's cells express "self" antigens. These antigens are different from those on the surface of bacteria or on the surface of virus-infected host cells ("non-self" or "foreign" antigens). The adaptive immune response is triggered by recognizing foreign antigen in the cellular context of an activated dendritic cell.

With the exception of non-nucleated cells (including erythrocytes), all cells are capable of presenting antigen through the function of major histocompatibility complex (MHC) molecules. Some cells are specially equipped to present antigen, and to prime naive T cells. Dendritic cells, B-cells, and macrophages are equipped with special "co-stimulatory" ligands recognized by co-stimulatory receptors on T cells, and are termed professional antigen-presenting cells (APCs).

Several T cells subgroups can be activated by professional APCs, and each type of T cell is specially equipped to deal with each unique toxin or microbial pathogen. The type of T cell activated, and the type of response generated, depends, in part, on the context in which the APC first encountered the antigen.

C. Autoimmune Treatments

In one aspect, the inventors contemplate treating individuals having been identified as having a greater risk of developing an immune disorder. In general, such treatments would involve administration of an agent that is anti-inflammatory, such as an NSAID, a corticosteroid, or other immune-suppressing agent, including biologicals such as toxilizumab, rituximab, ofatumumab, belimumab, epratuzumab, abatacept, golimumab, certolizumab, sifalimumab, i.v. immunoglobulin, anakinra, canakinumab and riolonocept.

V. Cancer

In some embodiments, the present disclosure relates to methods for predicting, diagnosing and/or treating cancer. Exemplary solid tumors can include, but are not limited to, a tumor of an organ selected from the group consisting of pancreas, colon, cecum, stomach, brain, head, neck, ovary, kidney, larynx, sarcoma, lung, bladder, melanoma, prostate, and breast. Exemplary hematological tumors include tumors of the bone marrow, T or B cell malignancies, leukemias, lymphomas, blastomas, myelomas, and the like. Further examples of cancers that may be treated using the methods provided herein include, but are not limited to, lung cancer (including small-cell lung cancer, non-small cell lung cancer, adenocarcinoma of the lung, and squamous carcinoma of the lung), cancer of the peritoneum, gastric or stomach cancer (including gastrointestinal cancer and gastrointestinal stromal cancer), pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, breast cancer, colon cancer, colorectal cancer, endometrial or uterine carcinoma, salivary gland carcinoma, kidney or renal cancer, prostate cancer, vulval cancer, thyroid cancer, various types of head and neck cancer, and melanoma.

The cancer may specifically be of the following histological type, though it is not limited to these: neoplasm, malignant; carcinoma; carcinoma, undifferentiated; giant and spindle cell carcinoma; small cell carcinoma; papillary carcinoma; squamous cell carcinoma; lymphoepithelial carcinoma; basal cell carcinoma; pilomatrix carcinoma; transitional cell carcinoma; papillary transitional cell carcinoma; adenocarcinoma; gastrinoma, malignant; cholangiocarcinoma; hepatocellular carcinoma; combined hepatocellular carcinoma and

cholangiocarcinoma; trabecular adenocarcinoma; adenoid cystic carcinoma; adenocarcinoma in adenomatous polyp; adenocarcinoma, familial polyposis coli; solid carcinoma; carcinoid tumor, malignant; branchiolo-alveolar adenocarcinoma; papillary adenocarcinoma; chromophobe carcinoma; acidophil carcinoma; oxyphilic adenocarcinoma; basophil carcinoma; clear cell adenocarcinoma; granular cell carcinoma; follicular adenocarcinoma; papillary and follicular adenocarcinoma; nonencapsulating sclerosing carcinoma; adrenal cortical carcinoma; endometroid carcinoma; skin appendage carcinoma; apocrine adenocarcinoma; sebaceous adenocarcinoma; ceruminous adenocarcinoma; mucoepidermoid carcinoma; cystadenocarcinoma; papillary cystadenocarcinoma; papillary serous cystadenocarcinoma; mucinous cystadenocarcinoma; mucinous adenocarcinoma; signet ring cell carcinoma; infiltrating duct carcinoma; medullary carcinoma; lobular carcinoma; inflammatory carcinoma; paget's disease, mammary; acinar cell carcinoma; adenosquamous carcinoma; adenocarcinoma w/squamous metaplasia; thymoma, malignant; ovarian stromal tumor, malignant; thecoma, malignant; granulosa cell tumor, malignant; androblastoma, malignant; sertoli cell carcinoma; leydig cell tumor, malignant; lipid cell tumor, malignant; paraganglioma, malignant; extra-mammary paraganglioma, malignant; pheochromocytoma; glomangiosarcoma; malignant melanoma; amelanotic melanoma; superficial spreading melanoma; lentigo malignant melanoma; acral lentiginous melanomas; nodular melanomas; malignant melanoma in giant pigmented nevus; epithelioid cell melanoma; blue nevus, malignant; sarcoma; fibrosarcoma; fibrous histiocytoma, malignant; myxosarcoma; liposarcoma; leiomyosarcoma; rhabdomyosarcoma; embryonal rhabdomyosarcoma; alveolar rhabdomyosarcoma; stromal sarcoma; mixed tumor, malignant; mullerian mixed tumor; nephroblastoma; hepatoblastoma; carcinosarcoma; mesenchymoma, malignant; brenner tumor, malignant; phyllodes tumor, malignant; synovial sarcoma; mesothelioma, malignant; dysgerminoma; embryonal carcinoma; teratoma, malignant; struma ovarii, malignant; choriocarcinoma; mesonephroma, malignant; hemangiosarcoma; hemangioendothelioma, malignant; kaposi's sarcoma; hemangiopericytoma, malignant; lymphangiosarcoma; osteosarcoma; juxtacortical osteosarcoma; chondrosarcoma; chondroblastoma, malignant; mesenchymal chondrosarcoma; giant cell tumor of bone; ewing's sarcoma; odontogenic tumor, malignant; ameloblastic odontosarcoma; ameloblastoma, malignant; ameloblastic fibrosarcoma; pinealoma, malignant; chordoma; glioma, malignant; ependymoma; astrocytoma; protoplasmic astrocytoma; fibrillary astrocytoma; astroblastoma; glioblastoma; oligodendroglioma; oligodendroblastoma; primitive neuroectodermal; cerebellar sarcoma; ganglioneuroblastoma; neuroblastoma; retinoblastoma; olfactory neurogenic tumor;

meningioma, malignant; neurofibrosarcoma; neurilemmoma, malignant; granular cell tumor, malignant; malignant lymphoma; Hodgkin's disease; hodgkin's; paragranuloma; malignant lymphoma, small lymphocytic; malignant lymphoma, large cell, diffuse; malignant lymphoma, follicular; mycosis fungoides; other specified non-Hodgkin's lymphomas; B-cell lymphoma; low grade/follicular non-Hodgkin's lymphoma (NHL); small lymphocytic (SL) NHL; intermediate grade/follicular NHL; intermediate grade diffuse NHL; high grade immunoblastic NHL; high grade lymphoblastic NHL; high grade small non-cleaved cell NHL; bulky disease NHL; mantle cell lymphoma; AIDS-related lymphoma; Waldenstrom's macroglobulinemia; malignant histiocytosis; multiple myeloma; mast cell sarcoma; immunoproliferative small intestinal disease; leukemia; lymphoid leukemia; plasma cell leukemia; erythroleukemia; lymphosarcoma cell leukemia; myeloid leukemia; basophilic leukemia; eosinophilic leukemia; monocytic leukemia; mast cell leukemia; megakaryoblastic leukemia; myeloid sarcoma; hairy cell leukemia; chronic lymphocytic leukemia (CLL); acute lymphoblastic leukemia (ALL); acute myeloid leukemia (AML); chronic myeloblastic leukemia (CML); and blastic plasmacytoid dendritic cell neoplasm (BPDCN).

A. Breast Cancer

Breast cancer is a cancer that starts in the breast, usually in the inner lining of the milk ducts or lobules. There are different types of breast cancer, with different stages (spread), aggressiveness, and genetic makeup. With best treatment, 10-year disease-free survival varies from 98% to 10%. Treatment is selected from surgery, drugs (chemotherapy), and radiation. In the United States, there were 216,000 cases of invasive breast cancer and 40,000 deaths in 2004. Worldwide, breast cancer is the second most common type of cancer after lung cancer (10.4% of all cancer incidence, both sexes counted) and the fifth most common cause of cancer death. In 2004, breast cancer caused 519,000 deaths worldwide (7% of cancer deaths; almost 1% of all deaths). Breast cancer is about 100 times as frequent among women as among men, but survival rates are equal in both sexes.

The first symptom, or subjective sign, of breast cancer is typically a lump that feels different from the surrounding breast tissue. According to the *The Merck Manual*, more than 80% of breast cancer cases are discovered when the woman feels a lump. According to the American Cancer Society, the first medical sign, or objective indication of breast cancer as detected by a physician, is discovered by mammogram. Lumps found in lymph nodes located in the armpits can also indicate breast cancer. Indications of breast cancer other than a lump may include changes in breast size or shape, skin dimpling, nipple inversion, or spontaneous

single-nipple discharge. Pain ("mastodynia") is an unreliable tool in determining the presence or absence of breast cancer, but may be indicative of other breast health issues.

When breast cancer cells invade the dermal lymphatics—small lymph vessels in the skin of the breast—its presentation can resemble skin inflammation and thus is known as inflammatory breast cancer (IBC). Symptoms of inflammatory breast cancer include pain, swelling, warmth and redness throughout the breast, as well as an orange-peel texture to the skin referred to as "peau d'orange." Another reported symptom complex of breast cancer is Paget's disease of the breast. This syndrome presents as eczematoid skin changes such as redness and mild flaking of the nipple skin. As Paget's advances, symptoms may include tingling, itching, increased sensitivity, burning, and pain. There may also be discharge from the nipple. Approximately half of women diagnosed with Paget's also have a lump in the breast.

Occasionally, breast cancer presents as metastatic disease, that is, cancer that has spread beyond the original organ. Metastatic breast cancer will cause symptoms that depend on the location of metastasis. Common sites of metastasis include bone, liver, lung and brain. Unexplained weight loss can occasionally herald an occult breast cancer, as can symptoms of fevers or chills. Bone or joint pains can sometimes be manifestations of metastatic breast cancer, as can jaundice or neurological symptoms. These symptoms are "non-specific," meaning they can also be manifestations of many other illnesses.

The primary risk factors that have been identified are sex, age, childbearing, hormones, a high-fat diet, alcohol intake, obesity, and environmental factors such as tobacco use, radiation and shiftwork. No etiology is known for 95% of breast cancer cases, while approximately 5% of new breast cancers are attributable to hereditary syndromes. In particular, carriers of the breast cancer susceptibility genes, BRCA1 and BRCA2, are at a 30-40% increased risk for breast and ovarian cancer, depending on in which portion of the protein the mutation occurs. Experts believe that 95% of inherited breast cancer can be traced to one of these two genes. Hereditary breast cancers can take the form of a site-specific hereditary breast cancer – cancers affecting the breast only – or breast- ovarian and other cancer syndromes. Breast cancer can be inherited both from female and male relatives.

Breast cancer subtypes are typically categorized on an immunohistochemical basis. Subtype definitions are generall as follows:

normal (ER+, PR+, HER2+, cytokeratin 5/6+, and HER1+)

luminal A (ER+ and/or PR+, HER2–)

luminal B (ER+ and/or PR+, HER2+)
triple-negative (ER–, PR–, HER2–)
HER2+/ER– (ER–, PR–, and HER2+)
unclassified (ER–, PR–, HER2–, cytokeratin 5/6–, and HER1–)

In the case of triple-negative breast cancer cells, the cancer's growth is not driven by estrogen or progesterone, or by growth signals coming from the HER2 protein. By the same token, such cancer cells do not respond to hormonal therapy, such as tamoxifen or aromatase inhibitors, or therapies that target HER2 receptors, such as Herceptin®. About 10-20% of breast cancers are found to be triple-negative. It is important to identify these types of cancer so that one can avoid costly and toxic effects of therapies that are unlike to succeed, and to focus on treatments that can be used to treat triple-negative breast cancer. Like other forms of breast cancer, triple-negative breast cancer can be treated with surgery, radiation therapy, and/or chemotherapy. One particularly promosing approach is "neoadjuvant" therapy, where chemo- and/or radiotherapy is provided prior to surgery. Another drug therapy is the use of poly (ADP-ribose) polymerase, or PARP inhibitors.

While screening techniques discussed above are useful in determining the possibility of cancer, a further testing is necessary to confirm whether a lump detected on screening is cancer, as opposed to a benign alternative such as a simple cyst. In a clinical setting, breast cancer is commonly diagnosed using a "triple test" of clinical breast examination (breast examination by a trained medical practitioner), mammography, and fine needle aspiration cytology. Both mammography and clinical breast exam, also used for screening, can indicate an approximate likelihood that a lump is cancer, and may also identify any other lesions. Fine Needle Aspiration and Cytology (FNAC), performed as an outpatient procedure using local anesthetic, involves attempting to extract a small portion of fluid from the lump. Clear fluid makes the lump highly unlikely to be cancerous, but bloody fluid may be sent off for inspection under a microscope for cancerous cells. Together, these three tools can be used to diagnose breast cancer with a good degree of accuracy. Other options for biopsy include core biopsy, where a section of the breast lump is removed, and an excisional biopsy, where the entire lump is removed.

Breast cancer screening is an attempt to find cancer in otherwise healthy individuals. The most common screening method for women is a combination of x-ray mammography and clinical breast exam. In women at higher than normal risk, such as those with a strong

family history of cancer, additional tools may include genetic testing or breast Magnetic Resonance Imaging.

Breast self-examination was a form of screening that was heavily advocated in the past, but has since fallen into disfavour since several large studies have shown that it does not have a survival benefit for women and often causes considerably anxiety. This is thought to be because cancers that could be detected tended to be at a relatively advanced stage already, whereas other methods push to identify the cancer at an earlier stage where curative treatment is more often possible.

X-ray mammography uses x-rays to examine the breast for any uncharacteristic masses or lumps. Regular mammograms are recommended in several countries in women over a certain age as a screening tool.

Genetic testing for breast cancer typically involves testing for mutations in the BRCA genes. This is not generally a recommended technique except for those at elevated risk for breast cancer.

The mainstay of breast cancer treatment is surgery when the tumor is localized, with possible adjuvant hormonal therapy (with tamoxifen or an aromatase inhibitor), chemotherapy, and/or radiotherapy. At present, the treatment recommendations after surgery (adjuvant therapy) follow a pattern. Depending on clinical criteria (age, type of cancer, size, metastasis) patients are roughly divided into high risk and low risk cases, with each risk category following different rules for therapy. Treatment possibilities include radiation therapy, chemotherapy, hormone therapy, and immune therapy.

Targeted cancer therapies are treatments that target specific characteristics of cancer cells, such as a protein that allows the cancer cells to grow in a rapid or abnormal way. Targeted therapies are generally less likely than chemotherapy to harm normal, healthy cells. Some targeted therapies are antibodies that work like the antibodies made naturally by one's immune system. These types of targeted therapies are sometimes called immune-targeted therapies.

There are currently 3 targeted therapies doctors use to treat breast cancer. Herceptin® (trastuzumab) works against HER2-positive breast cancers by blocking the ability of the cancer cells to receive chemical signals that tell the cells to grow. Tykerb® (lapatinib) works against HER2-positive breast cancers by blocking certain proteins that can cause uncontrolled cell growth. Avastin® (bevacizumab) works by blocking the growth of new blood vessels that cancer cells depend on to grow and function.

Hormonal (anti-estrogen) therapy works against hormone-receptor-positive breast cancer in two ways: first, by lowering the amount of the hormone estrogen in the body, and second, by blocking the action of estrogen in the body. Most of the estrogen in women's bodies is made by the ovaries. Estrogen makes hormone-receptor-positive breast cancers grow. So reducing the amount of estrogen or blocking its action can help shrink hormone-receptor-positive breast cancers and reduce the risk of hormone-receptor-positive breast cancers coming back (recurring). Hormonal therapy medicines are not effective against hormone-receptor-negative breast cancers.

There are several types of hormonal therapy medicines, including aromatase inhibitors, selective estrogen receptor modulators, and estrogen receptor downregulators. In some cases, the ovaries and fallopian tubes may be surgically removed to treat hormone-receptor-positive breast cancer or as a preventive measure for women at very high risk of breast cancer. The ovaries also may be shut down temporarily using medication.

In planning treatment, doctors can also use PCR tests like Oncotype DX or microarray tests that predict breast cancer recurrence risk based on gene expression. In February 2007, the first breast cancer predictor test won formal approval from the Food and Drug Administration. This is a new gene test to help predict whether women with early-stage breast cancer will relapse in 5 or 10 years, this could help influence how aggressively the initial tumor is treated.

Radiation therapy is also used to help destroy cancer cells that may linger after surgery. Radiation can reduce the risk of recurrence by 50-66% when delivered in the correct dose.

B. Colorectal Cancer

Colorectal cancer (CRC), also known as bowel cancer and colon cancer, is the development of cancer from the colon or rectum (parts of the large intestine). A cancer is the abnormal growth of cells that have the ability to invade or spread to other parts of the body. Signs and symptoms may include blood in the stool, a change in bowel movements, weight loss, and feeling tired all the time.

Most colorectal cancers are due to old age and lifestyle factors with only a small number of cases due to underlying genetic disorders. Some risk factors include diet, obesity, smoking, and lack of physical activity. Dietary factors that increase the risk include red and processed meat as well as alcohol. Another risk factor is inflammatory bowel disease, which includes Crohn's disease and ulcerative colitis. Some of the inherited genetic disorders that

can cause colorectal cancer include familial adenomatous polyposis and hereditary non-polyposis colon cancer; however, these represent less than 5% of cases. It typically starts as a benign tumor, often in the form of a polyp, which over time becomes cancerous.

Bowel cancer may be diagnosed by obtaining a sample of the colon during a sigmoidoscopy or colonoscopy. This is then followed by medical imaging to determine if the disease has spread. Screening is effective for preventing and decreasing deaths from colorectal cancer. Screening, by one of a number of methods, is recommended starting from the age of 50 to 75. During colonoscopy, small polyps may be removed if found. If a large polyp or tumor is found, a biopsy may be performed to check if it is cancerous. Aspirin and other non-steroidal anti-inflammatory drugs decrease the risk. Their general use is not recommended for this purpose, however, due to side effects.

Treatments used for colorectal cancer may include some combination of surgery, radiation therapy, chemotherapy and targeted therapy. Cancers that are confined within the wall of the colon may be curable with surgery while cancer that has spread widely are usually not curable, with management being directed towards improving quality of life and symptoms. The five-year survival rate in the United States is around 65%. The individual likelihood of survival depends on how advanced the cancer is, whether or not all the cancer can be removed with surgery, and the person's overall health. Globally, colorectal cancer is the third most common type of cancer, making up about 10% of all cases. In 2012, there were 1.4 million new cases and 694,000 deaths from the disease. It is more common in developed countries, where more than 65% of cases are found. It is less common in women than men.

The signs and symptoms of colorectal cancer depend on the location of the tumor in the bowel, and whether it has spread elsewhere in the body (metastasis). The classic warning signs include: worsening constipation, blood in the stool, decrease in stool caliber (thickness), loss of appetite, loss of weight, and nausea or vomiting in someone over 50 years old. While rectal bleeding or anemia are high-risk features in those over the age of 50, other commonly described symptoms including weight loss and change in bowel habit are typically only concerning if associated with bleeding.

Greater than 75–95% of colorectal cancer occurs in people with little or no genetic risk. Risk factors include older age, male gender, high intake of fat, alcohol, red meat, processed meats, obesity, smoking, and a lack of physical exercise. Approximately 10% of cases are linked to insufficient activity. The risk from alcohol appears to increase at greater than one drink per day. Drinking 5 glasses of water a day is linked to a decrease in the risk of colorectal cancer and adenomatous polyps. *Streptococcus gallolyticus* is associated with

colorectal cancer. Some strains of *Streptococcus bovis/Streptococcus equinus* complex are consumed by millions of people daily and thus may be safe. 25 to 80% of people with *Streptococcus bovis/gallolyticus* bacteremia have concomitant colorectal tumors. Seroprevalence of *Streptococcus bovis/gallolyticus* is considered as a candidate practical marker for the early prediction of an underlying bowel lesion at high risk population. It has been suggested that the presence of antibodies to *Streptococcus bovis/gallolyticus* antigens or the antigens themselves in the bloodstream may act as markers for the carcinogenesis in the colon.

Colorectal cancer diagnosis is performed by sampling of areas of the colon suspicious for possible tumor development, typically during colonoscopy or sigmoidoscopy, depending on the location of the lesion. It is confirmed by microscopical examination of a tissue sample.

Disease extent is usually determined by a CT scan of the chest, abdomen and pelvis. Other potential imaging tests such as PET and MRI may be used in certain cases.

Colon cancer staging is done next and is based on radiology and pathology. As for all other forms of cancer, tumor staging is based on the TNM system which considers how much the initial tumor has spread, if and where there are lymph node metastasis and if there are metastases in more distant organs, usually liver.

The microscopic cellular characteristics of the tumor are reported from the analysis of tissue taken from a biopsy or surgery. A pathology report contains a description of the microscopical characteristics of the tumor tissue, including both tumor cells and how the tumor invades into healthy tissues and finally if the tumor appears to be completely removed. The most common form of colon cancer is adenocarcinoma. Other, rarer types include lymphoma, adenosquamous and squamous cell carcinoma. Some subtypes have been found to be more aggressive.

The treatment of colorectal cancer can be aimed at cure or palliation. The decision on which aim to adopt depends on various factors, including the person's health and preferences, as well as the stage of the tumor. When colorectal cancer is caught early, surgery can be curative. However, when it is detected at later stages (for which metastases are present), this is less likely and treatment is often directed at palliation, to relieve symptoms caused by the tumor and keep the person as comfortable as possible.

If the cancer is found at a very early stage, it may be removed during a colonoscopy. For people with localized cancer, the preferred treatment is complete surgical removal with adequate margins, with the attempt of achieving a cure. This can either be done by an open

laparotomy or sometimes laparoscopically. The colon may then be reconnected or a person may have a colostomy.

If there are only a few metastases in the liver or lungs they may also be removed. Sometimes chemotherapy is used before surgery to shrink the cancer before attempting to remove it. The two most common sites of recurrence of colorectal cancer are the liver and lungs.

In both cancer of the colon and rectum, chemotherapy may be used in addition to surgery in certain cases. The decision to add chemotherapy in management of colon and rectal cancer depends on the stage of the disease.

In Stage I colon cancer, no chemotherapy is offered, and surgery is the definitive treatment. The role of chemotherapy in Stage II colon cancer is debatable, and is usually not offered unless risk factors such as T4 tumor or inadequate lymph node sampling is identified. It is also known that the people who carry abnormalities of the mismatch repair genes do not benefit from chemotherapy. For stage III and Stage IV colon cancer, chemotherapy is an integral part of treatment.

If cancer has spread to the lymph nodes or distant organs, which is the case with stage III and stage IV colon cancer respectively, adding chemotherapy agents fluorouracil, capecitabine or oxaliplatin increases life expectancy. If the lymph nodes do not contain cancer, the benefits of chemotherapy are controversial. If the cancer is widely metastatic or unresectable, treatment is then palliative. Typically in this setting, a number of different chemotherapy medications may be used. Chemotherapy drugs for this condition may include capecitabine, fluorouracil, irinotecan, oxaliplatin and UFT. The drugs capecitabine and fluorouracil are interchangeable, with capecitabine being an oral medication while fluorouracil being an intravenous medicine. Some specific regimens used for CRC are FOLFOX, FOLFOXIRI, and FOLFIRI. Antiangiogenic drugs such as bevacizumab are often added in first line therapy. Another class of drugs used in the second line setting are epidermal growth factor receptor inhibitors, of which the two FDA approved ones are cetuximab and panitumumab.

The primary difference in the approach to low stage rectal cancer is the incorporation of radiation therapy. Often, it is used in conjunction with chemotherapy in a neoadjuvant fashion to enable surgical resection, so that ultimately as colostomy is not required. However, it may not be possible in low lying tumors, in which case, a permanent colostomy may be required. Stage IV rectal cancer is treated similar to stage IV colon cancer.

While a combination of radiation and chemotherapy may be useful for rectal cancer, its use in colon cancer is not routine due to the sensitivity of the bowels to radiation. Just as for chemotherapy, radiotherapy can be used in the neoadjuvant and adjuvant setting for some stages of rectal cancer.

Immunotherapy with immune checkpoint inhibitors has been found to be useful for a type of colorectal cancer with mismatch repair deficiency and microsatellite instability. Most people who do improve, however, still worsen after months or years. Other types of colorectal cancer as of 2017 is still being studied.

Palliative care is medical care which focuses on treatment of symptoms from serious illness, like cancer, and improving quality of life. Palliative care is recommended for any person who has advanced colon cancer or has significant symptoms. Involvement of palliative care may be beneficial to improve the quality of life for both the person and his or her family, by improving symptoms, anxiety and preventing admissions to the hospital.

In people with incurable colorectal cancer, palliative care can consist of procedures that relieve symptoms or complications from the cancer but do not attempt to cure the underlying cancer, thereby improving quality of life. Surgical options may include non-curative surgical removal of some of the cancer tissue, bypassing part of the intestines, or stent placement. These procedures can be considered to improve symptoms and reduce complications such as bleeding from the tumor, abdominal pain and intestinal obstruction. Non-operative methods of symptomatic treatment include radiation therapy to decrease tumor size as well as pain medications.

VI. Examples

The following examples are included to demonstrate preferred embodiments of the disclosure. It should be appreciated by those of skill in the art that the techniques disclosed in the examples which follow represent techniques discovered by the inventor to function well in the practice of the disclosure, and thus can be considered to constitute preferred modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the disclosure.

EXAMPLE 1 - Methods

The BCR heavy chain repertoires used in this study were collected in the context of other studies and shared with us upon request ([7] and under review with *Multiple Sclerosis Journal*). The samples from each study are referred to by the year the study was completed (Table 1). The repertoires were obtained as described in [7]. Briefly, CSF was taken by lumbar puncture from patients diagnosed with either RRMS or OND. DNA was extracted from CSF cell pellets, and targeted PCR was conducted to amplify rearranged BCR genes. Because of the limited amount of DNA extracted from each CSF sample, the PCR amplification protocol has been designed and carefully implemented to minimize the impact of amplification bias and carry-over contamination between samples. DNA is first amplified using whole genome amplification followed by targeted PCR amplification of the variable region of BCR heavy chain genes utilizing a V gene segment from the VH4 family. VH4 sequences were targeted because previous studies found VH4 usage to be elevated in patients with RRMS [15, 16]. Sequencing was conducted on the 454 platform. All tissue and patient data from both studies was handled in accordance with IRB-approved protocols.

The DNA sequences for each sample were processed to prepare them for analysis following recommendations in [17]. Specifically, sequences with a length less than 300 base pairs or an average quality score less than 35 were removed. The regions of each sequence to which the PCR primers hybridize were trimmed, and duplicate sequences appearing within a single sample were counted and then collapsed to a single sequence. The remaining sequences were aligned to a database of germline gene segments for V, D, and J gene assignment. Sequences representing non-functional rearrangements were removed. Processing was performed using the pRESTO [18], IgBlast [19], and RepCalc pipelines on the VDJServer Immune Repertoire Analysis Portal (world-wide-web at vdjserver.org).

For the sequences remaining after processing, the CDR3 nucleotide sequences were identified, according to the Immunogenetics Information System (world-wide-web at imgt.org) definitions. CDR3 sequences containing ambiguous base calls were removed, and the remaining sequences were compared across samples to identify potential carry-over contamination. The amount observed was in line with other studies [20, 21]. Sequences observed in more than one sample were removed. The remaining CDR3 sequences were used as input to develop the statistical classifier as described above.

EXAMPLE 2 - Results

The inventors' overall approach was as follows. They used two data sets, one as training data and one as validation data (Table 1). The training data set was used with exhaustive leave-one-out cross-validation for model selection to identify the best model from among seven models tested (Table 2). The seven models correspond to different approaches to representing immune receptor sequences. The model with highest classification accuracy by cross-validation was selected for application to the validation data set.

The training data set consisted of 23 patients, 11 with RRMS and 12 with OND (2015 Study, Table 1). The validation data set consisted of 102 patients, 60 with RRMS and 42 with OND (2017 Study, Table 1). For both studies, B cell repertoires were collected and processed as described in [7]. Briefly, samples were collected from patient cerebrospinal fluid (CSF) (FIG. 1A), and VH4-containing BCR heavy chain genes were sequenced using next generation sequencing (FIG. 1B). VH4-containing heavy chains were targeted because previous studies found elevated VH4 expression in patients with RRMS [2, 7]. Sequence pre-processing was performed as described in the Methods to identify CDR3 sequences for input into the method.

The inventors utilized the CDR3 sequence of each heavy chain gene, because it is the somatically generated portion of the gene and the primary determinant of the antigen binding specificity encoded by the gene. To accommodate the varying length of CDR3, each CDR3 sequence was cut into snippets of equal length (*i.e.*, k-mers). The inventors considered snippet lengths of 2, 4, 5, 6, and 7 amino acids or codons. For each CDR3, the full set of overlapping snippets was used. The inventors considered three different sequence representations: DNA sequence, amino acid sequence, and a representation based on Atchley factors (FIG. 1C). There are five Atchley factors derived from a set of over 50 amino acid properties by dimensionality reduction to identify clusters of amino acid properties that co-vary [10]. The five Atchley factors correspond loosely to polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. For the Atchley factor representation, each amino acid in a snippet is represented by a vector of its five Atchley factor values. The inventors conducted model selection over seven combinations of snippet length and sequence representation (Table 2).

The approach was applied the to a training data set of 23 patients using one-holdout cross-validation (FIG. 2A). Classification accuracy on the holdout patients was used to identify the best performing model from among the seven models tried (Table 2). A snippet

size of 6 amino acid residues resulted in the highest classification accuracy. Categorical representations of the DNA nucleotides and amino acid residues both underperformed the Atchley factor representation. The best performing model correctly diagnosed 20 out of 23 patients (Table 2, FIG. 3A). A plot of the ROC curve for the best model shows the true positive versus false positive rate as a function of the threshold required to obtain a positive diagnosis (FIG. 3B). The area under the curve is 0.86.

To determine if the best performing model generalizes to unused data, the full 23-patient training set was used to fit the weights and bias term, and the resulting model was applied to score a validation data set of 102 patients (FIG. 2B). The model correctly diagnoses 73 out of 102 patients, corresponding to an accuracy of 72% (FIG. 3C). The ROC curve for the validation data is shown in FIG. 3d. The area under the curve is 0.75.

To discern the biochemical features of snippets flagging a positive diagnosis, the inventors examined the weights of the best performing model with parameters fit on the full 23-patient training set. The weights reveal the relative importance of each Atchley factor along every position of the snippet (FIG. 4). These weights, together with the Atchley factor values, form a biochemical motif that indicates an RRMS diagnosis. The inventors observe relatively large, negative weights along almost every position of the snippet for Atchley factors II and IV, indicating a high probability of an RRMS diagnosis for snippets with negative values for these two Atchley factors. In particular, they notice large negative weights for factor II for positions 1 and 5 and for factor IV for positions 1, 3, and 4. A negative value for Atchley factor II correlates with amino acid residues that appear frequently in α-helical segments. A negative value for Atchley factor IV correlates with amino acid residues less commonly used and having high heat capacity and refractivity. The weights for the other Atchley factors are position-dependent. The inventors observe relatively large positive weights for Atchley factor I at position 1 and for Atchley factor V at position 3. They also observe relatively large negative weights for positions 1 and 3 for Atchley factor III. This indicates increased probability of an RRMS diagnosis for snippets with large, positively charged, hydrophilic residues at snippet positions 1 and 3.

The inventors next aligned the highest scoring snippet from each patient to determine where within CDR3 the diagnostic snippet is positioned (FIG. 5). They find that the highest scoring snippets can be located anywhere along CDR3. Although the snippet sequences do not align well, patterns are observable in their Atchley factors, which are shown next to each snippet (FIG. 5). Consistent with the values for the weights, they observe a tendency toward

hydrophilicty for snippet position 1, toward α-helical values at position 5, toward high heat capacity and refractivity at positions 1 through 4, and toward negative charge at position 6.

The inventors next looked at the distribution of snippet scores in the 23 patients of the training data set (FIG. 6). Only 27 of 3259 snippets score above 0.5 (the threshold for a RRMS diagnosis), and all of these were from RRMS patient repertoires. Each RRMS patient had no more than 5 snippets that scored above the threshold.

To determine if the rarity of high scoring snippets in the patient repertoires can be attributed to the likelihood of the corresponding DNA sequences arising by chance in V(D)J recombination junctions, the inventors examined the DNA encodings of each snippet. For each amino acid sequence, there are many possible DNA encodings. An example of how to calculate the total number of encodings for a single snippet is shown in FIG. 7. They find that the diagnostic snippets identified by the model have significantly fewer possible encodings than non-diagnostic snippets (p-value is 7.41×10^{-8}). Under the naïve assumption that CDR3 sequence is generated at random, RRMS diagnostic snippets would be some of the least likely to occur.

EXAMPLE 3 - DISCUSSION

High-throughput sequencing of immune repertoires now enables their detailed characterization, driving interest in utilizing repertoires in clinical applications, including diagnosing and prognosticating diseases (*e.g.*, [12]). Attempts to date have taken the approach of computing repertoire-level summary statistics, such as gene segment usage statistics, repertoire diversity, and clonality, and looking for differences in these statistics between two sets of repertoires (*e.g.*, cases and controls) [1-7]. This approach captures important features of a repertoire as a whole, and it can give insight into the biological processes underlying repertoire differences, such as whether there is clonal expansion or recruitment of new cells. On the other hand, this approach ignores the vast amount of information available in the individual immune receptor sequences, in particular, information about the encoded antigen binding specificities.

The inventors present here a new approach that allows application of standard machine learning techniques to mine the full set of repertoire sequences for sequence patterns that distinguish one group of repertoires from the other. There are two key features of this approach. The first is that one captures all k-mers from all CDR3s in a repertoire and represents them as biochemical features using Atchley factors. The second is that one scores

all k-mers in a repertoire and then aggregates the set of scores to predict a label for the whole repertoire.

The inventors have focused on the CDR3 portion of immune receptor sequences, because it is the somatically generated portion of the gene and the primary determinant of the antigen binding specificity encoded by the gene. The approach could be readily applied, however, to other parts of the gene, and even to the full sequence. The longer the sequence used, however, the more training data would be required to accommodate the corresponding increase in the number of model parameters.

To accommodate variation in CDR3 length, the inventors represent each CDR3 sequence as a set of overlapping k-mers of a specified length. For example, a CDR3 of eight amino acids in length would be represented as three 6-mers. In their MS application, the inventors used k-mers varying from four to seven amino acids and found that the highest classification accuracy was achieved with 6-mers (Table 2). Some CDR3s in these data sets are only six amino acids in length and are thus excluded from analysis when longer k-mers are used. Shorter k-mers, on the other hand, are more likely to appear in both MS and OND repertoires and are therefore not useful for discrimination if the relative abundance of the receptor is not utilized.

The inventors hypothesized that using a biochemical representation of amino acid k-mers would be beneficial, because such a representation captures sequence features related to receptor-antigen binding, and therefore related to the receptor's function. Additionally, immune receptors with distinct amino acid sequences that bind to the same antigen would be expected to have similar biochemical properties. Indeed, the inventors found that, for a fixed k-mer length of six amino acids, the biochemical representation resulted in higher classification accuracy than either an amino acid or DNA sequence representation (Table 2).

To aggregate the scores from all k-mers in a repertoire to a repertoire-level label, the inventors took the maximum score based on the assumption that, among all receptors in a repertoire, those participating in the phenotype-related immune response may be rare. Thus, the inventors wanted a function that would flag a positive diagnosis even for a single high-scoring snippet. The maximum score is a special case of the generalized mean, however, and other means, or even other functions, could be used to accommodate different assumptions about the underlying immune response and its role in the phenotype.

Using this approach, the inventors were able to mine the individual CDR3 sequences of OND and RRMS patient repertoires to discover a biochemical motif that correctly classifies repertoires according to diagnosis with accuracy of 87% on training data and 72%

on validation data. Importantly, no prior knowledge of the disease was utilized (*i.e.*, it was not necessary to know which antigens the B cells may be responding to). Additionally, the method did not rely on finding "public clones," as the inventors removed all shared sequences to control for possible carry-over contamination, as described in Methods.

In the context of MS, a classification accuracy of 72% is highly significant. MS is an autoimmune disease that is difficult to diagnose. There are no single symptoms, physical findings, or laboratory tests that provide a definitive MS diagnosis [13]. The current method of diagnosis relies on the 2010 revisions to the McDonald criteria and requires demonstration of dissemination of central nervous system lesions in both space and time, along with the exclusion of other diagnoses [8]. Currently, the most widely used piece of paraclinical evidence for MS diagnosis is magnetic resonance imaging (MRI). Therefore, the accuracy obtained using MRI to distinguish patients with MS from those with OND is the most appropriate direct comparison for the observed 72% accuracy. The inventors know of one study based on the most recent MRI criteria making this assessment. An accuracy of 57% was observed for distinguishing MS from primary and secondary central nervous system vasculitis, lupus, and Sjogren's syndrome [14]. In this context, a classification accuracy of 72% is highly significant.

Table 1 - Repertoire Sequencing Data Sets Used to Develop and Test the MS Classifier*

	Relapsing Remitting Multiple Sclerosis	**Other Neurological Disease**
2015 Study [7]	11	12
2017 Study	60	42

* The number of patients in each study with each diagnosis is shown.

Table 2 - Sequence Representations Used for Model Selection*

Snippet Length	**Sequence Representation**	**Classification Accuracy on the Training Data Set by Exhaustive 1-Holdout Cross-Validation***
4 Amino Acids	Atchley Factors	11/23
5 Amino Acids	Atchley Factors	15/23

6 Amino Acids	Atchley Factors	20/23
7 Amino Acids	Atchley Factors	14/23
2 DNA Triples	DNA Nucleotides	12/23
6 DNA Triplets	DNA Nucleotides	8/23
6 Amino Acids	Amino Acid Residue	15/23

* Results reported as the fraction of cases where the model's best guess on the diagnosis is correct; CDR3 sequences were cut into snippets of varying length and represented as DNA sequence, amino acid sequence, or Atchley factors [10].

EXAMPLE 4 – CANCER METHODS

T-cell receptor datasets for colorectal and breast cancer. The inventors searched for existing datasets of T-cell receptor (TCR) sequences extracted from tumor biopsies and healthy matching control tissue. The inventors found data from 14 colorectal cancer patients published by Sherwood *et al.* and data from 16 breast cancer patients published by Beausang *et al.* [24, 25]. In both studies, adjacent healthy control tissue was biopsied from each cancer patient at the same time as their tumor, providing patient-matched control tissue for each tumor sample. All immune receptor sequencing was done by Adaptive Biotechnologies. Data from the two studies is summarized in Tables 3 and 4.

The inventors used the data to fit their statistical classifier to categorize a TCR repertoire as deriving either from tumor or healthy tissue. They treated each cancer type separately, resulting in one model for colorectal cancer and another model for breast cancer. The performances of the statistical classifiers can be assessed by a patient-holdout cross-validation, where both the tumor and healthy matching control repertoires from the same patient were held-out for validation.

Representing the specificity of the T-cell receptor CDR3 sequence. To determine how to profile the specificity of a T-cell receptor CDR3 sequence, the inventors analyzed X-ray crystallographic structures of human T-cell receptors bound to an antigen-MHC complex. Preliminary analysis revealed that CDR3 residues in direct contact with the antigen (≥ 5Å) tended to lie directly adjacent to each other along the CDR3 sequence, forming a contiguous strip (FIG. 8A). To verify this observation, the inventors extracted CDR3 sequences from 57

T-cell receptor structures and annotated each residue as either 'C' for being in direct contact with the antigen or 'Ø' for not being in contact with the antigen. The inventors used the annotations to perform a multiple sequence alignment, forcing the aligner to match contact positions together (FIG. 8B). The alignment confirmed that contact residues tend to form a nearly contiguous strip of residues. While the size and relative location of this strip varied with each T-cell receptor, no additional regions in the CDR3 appeared to directly touch the antigen. The average length of the strip is 4 and the strip rarely appears in the first 3 or last 3 residues of the CDR3.

To capture the residues directly in contact with an antigen, the inventors exclude the first and last 3 residues of the CDR3 and partition the remaining CDR3 sequence into every possible contiguous strip of 4 amino acid residues (FIG. 8C). The hypothesis is that one of these snippets directly contacts the receptor's cognate antigen, although the correct one is not known. The challenge is to identify the snippet in contact with a tumor antigen within the pool of CDR3 sequences from a tumor biopsy.

To identify snippets that have different amino acid sequences but that may bind the same or similar antigens, the inventors represent each snippet using biochemical Atchley factor values [23]. There are five Atchley factors derived from a set of over 50 amino acid properties by dimensionality reduction to identify clusters of amino acid properties that co-vary. The five Atchley factors correspond loosely to polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. For the Atchley factor representation, each amino acid in a snippet is represented by a vector of its five Atchley factor values (FIG. 8D).

Representing the abundance of each receptor. The number of identical T-cell receptors in a sample can reveal if a T-cell has undergone clonal expansion in response to an antigen. This makes it an important feature for the statistical classifier. There are two important considerations when using the receptor quantity as a feature. First, both the DNA yield and sequencing depth coverage affect the measurement. Therefore, the receptor quantity is only meaningful relative to the total number of receptors in a sample. For this reason, the relative abundance is used in place of the raw receptor count. The other important consideration is that a T-cell can proliferate at an exponential rate in response to its antigen. As a result, small differences in the affinity for an antigen can result in exponentially large differences in receptor quantity. Therefore, the inventors take the logarithm of the receptors' relative abundances. They hypothesize that the log-term better relates the quantity of a receptor to its affinity for its antigen, and they use this as an additional feature alongside the biochemical Atchley factor values.

The inventors considered two approaches for computing the relative abundance of the snippet given the relative abundances of the receptors. Each approach represents a different strategy for coping with identical snippets appearing in different CDR3 sequences. In the first method, the inventors identify every CDR3 sequence that contains a given snippet and the inventors calculate the relative abundance of the snippet by first summing the abundances of all the CDR3 sequences containing the snippet. This provides the total count of the snippet across the sample. They then divide by the total count of all snippets. The second approach is to treat every time a snippet appears in a different CDR3 sequence as an individual instance. For each instance, the relative abundance of the receptor containing that instance is used as the relative abundance of that instance of the snippet. For reasons outlined below, the inventors simply use the relative abundance of the most abundant CDR3 sequence containing that snippet. It is unclear to us which of the two approaches is better, so they assessed the performance of the statistical classifier under both approaches.

Normalizing the Features. It is important to normalize the features of a statistical classifier before they are used. Assuming an equal representation of all 20 amino acid residues, the inventors normalized the Atchley factor values so that each of the 5 biochemical descriptors has zero mean and unit variance. It is unclear whether it is appropriate to normalize the log-term of the relative abundance. Therefore, the inventors assess the performance of the statistical classifier with and without the log-term being normalized.

Scoring each sequence in a repertoire. Every snippet from every CDR3 sequence in a biopsy is scored by a detector function indicating if a snippet predicts the tissue is tumor. The inventors use a logistic function because of its widespread use and simplicity, and because it models the outcome of a two-category process. The first step is to compute a biased, weighted sum of the snippet's features, referred to as a logit.

$$\text{logit} = b_0 + W_1 \cdot f_1 + W_2 \cdot f_2 + \cdots + W_{20} \cdot f_{20} + W_{21} \cdot \ln q \quad (1)$$

The values f_1 through f_{20} represent the five Atchley factors from the four residues in the snippet. The value q represents the relative abundance. The bias term b_0 along with the weights W_1 through W_{21} are the parameters of the model and are fit by maximum likelihood using gradient descent optimization techniques as described below. The same weights W_1 through W_{21} and bias term b_0 are used for all snippets. Once the logit is computed, the value is passed through the sigmoid function to obtain a score between 0 and 1:

$$\text{score} = {}^{1}/_{1 + e^{-\text{logit}}} \quad (2)$$

Aggregation of snippet scores to predict tumor or healthy control tissue. A biopsy's snippet scores need to be aggregated into a single value to categorize a sample as tumor or healthy tissue. Because only a small fraction of the T-cells is expected to be responding to shared antigens across tumors, it is necessary to categorizing a sample as tumor in the presence of even just a small number of high scoring snippets. This is accomplished by categorizing a sample as tumor when even a single high scoring snippet is found. Assuming the output of the detector function represents a probability value between 0 and 1, the form of the model can be written as:

$$P(\text{ positive diagnosis} \mid \text{snip}_1, \text{snip}_2, \text{snip}_3, \ldots) = \text{Maximum}(\text{ score}_1, \text{score}_2, \text{score}_3, \ldots\ ) \quad (3)$$

The model predicts a tumor whenever even a single snippet has a score above 0.5, and predicts healthy tissue only when every snippet has a score below 0.5.

When the same snippet appears in different receptors and the inventors pair each instance of the snippet with the quantity of the corresponding receptor, they use only the receptor with the largest quantity and discard all other instances of the snippet. This works because only the highest score contributes to the model's predictions.

Parameter Fitting by Gradient Descent. Specific values for the weights W_1 through W_{21} and bias term b_0 in the detector function are determined using tissue information (*i.e.* tumor *vs* healthy). The values must be chosen to maximize the likelihood that each prediction is correct. To search for the optimal values, gradient optimization techniques are used. With these techniques, each parameter is iteratively adjusted along the gradient in a direction that maximizes the log-likelihood, which in turn maximizes the likelihood that each prediction is correct. The initial value for the bias term b_0 is 0, and initial values for the weights on the Atchley factors W_1 through W_{20} are drawn at random according to $W \sim \mathcal{N}(0, N_{\text{features}}^{-1} = 1/20)$. Different protocols for initializing weight W_{21} on the relative abundance term were tried as reported in Table 3. Because the Adam optimizer, a gradient descent-based method, has been shown to work well on a wide range of optimization tasks, it is used here. The Adam optimizer is run for 2500 iterations with a step size of 0.01. The default values for the other Adam optimizer settings are: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

A limitation of using a gradient descent-based method is there is no guarantee of finding the globally optimal solution. Although the chosen detector function constitutes a linear model, the scores from every snippet are aggregated together in a non-linear fashion. Multiple local minima could exist. To address this, 2.5×10^5 runs of Adam optimization,

each starting from different initial parameters $W_i \sim \mathcal{N}(0, N_{\text{features}}^{-1})$, are used, and the best fit solution over all runs is used to categorize new tissue samples. To spare computational resources, fewer runs were done on models that failed to perform well after a smaller number of runs.

EXAMPLE 5 – CANCER RESULTS

Development of the model and validation. The inventors applied the above-described approach to the training datasets for both colorectal and breast cancer. Each dataset is treated separately, resulting in one model for colorectal cancer and another model for breast cancer. To assess the performance of each model, the inventors performed a patient-holdout cross-validation, where the tumor and patient-matched healthy control tissue are simultaneously held out for validation (FIG. 9). The tumor and patient-matched healthy control tissue are scored independently, and the model has no knowledge that because one tissue it tumor then the other tissue must be healthy.

As described above, several variations of the model are considered. These include different methods for calculating the relative abundance, different approaches for normalizing the log-term as a feature, and different initialization schemes for weight W_{21} of the log-term. Results are reported in Table 5. Numerous other modifications are also reported in Table 5, most of which were tried only on the colorectal cancer dataset. The best performing models for colorectal and breast cancer are highlighted in red and are the sole focus for the rest of this study.

Colorectal Cancer Results. The majority of models tried on the colorectal data set performed better than baseline (a classification accuracy of 50%). The best model was obtained using the relative abundance of each snippet (rather than the receptor) with the log-term left un-normalized and its weight (W_{21}) initialized to 0. The model correctly categorizing $26/28 \approx 93\%$ of the samples and always scored the tumor above the patient matched healthy control tissue (FIG. 10A). That the model always scored the tumor above the patient matched healthy control tissue is significant because the model never had access to the information that the samples came from the same patient.

To discern the biochemical features of the snippets resulting in a tumor categorization, the inventors examined the weights of the model with parameters fit on all 14 patients in the training set. The weights reveal how each Atchley factor contributes to the score and the relative importance of each position (FIG. 10B). The inventors observe mostly

negative weights along almost every position of the snippet for Atchley factors II and IV, indicating a high probability of tumor for snippets that contain residues that participate in α-helical segments and that appear infrequently among the space of all protein sequences. The inventors also observe only positive weights for Atchley factor V, indicating a high probability of tumor for snippets enriched with residues with positively charged sidechains. Many of the remaining weights are position dependent. The weight on the log-term favors snippets that exhibit a large relative abundance, which is not surprising.

Next, the inventors aligned all snippets that scored high enough to categorize a sample as tumor and embedded these snippets back into their original CDR3 sequences (FIG. 10C). Whenever a snippet is found in multiple CDR3 sequences, they show the CDR3 sequence with the highest relative abundance. Next to each CDR3, the inventors list the sequence's rank based on its relative abundance. While several of the CDR3 sequences are highly enriched, many of the sequences are not. These CDR3 sequences would have been missed if they had simply examined the top clones.

Breast Cancer Results. As for the colorectal cancer data set, the inventors tried a variety of models on the breast cancer data set and found that the best performing model on breast cancer was the one using the relative abundance of the receptor rather than the snippet. The best model was otherwise similar to that for colorectal cancer with log-term left un-normalized and its weight (W_{21}) initialized to 0. The model correctly categorized $30/32 \approx 94\%$ of the samples and always scored the tumor above the patient-matched healthy control tissue (FIG. 11A).

Next, the inventors examined the weights of the model with parameters fit on all 16 patients in the training set (FIG. 11B). The direction and magnitude of the weights changed considerably from those obtained on the colorectal samples, indicating that the model is specific to the cancer type. It appears that for all Atchley factor values, the weights are position dependent. The inventors observe that for Atchley factors I and II, a negative weight at the first position in the snippet and a steady transition to a positive weight at the last position results in a snippet receiving a high score. The exact opposite trend is observed for Atchley factor IV, where the weights start off positive and steadily transition to very large negative weights. The one similarity with the colorectal results is that the model favors receptors with a high relative abundance, as observed in the weight on the log-term of the model.

The inventors aligned all snippets that scored high enough to categorize a sample as tumor and embedded these snippets back into their original CDR3 sequences (FIG. 11C).

Whenever a snippet is found on multiple CDR3 sequences, the inventors show the CDR3 sequence with the highest relative abundance. When they look at the rank of each CDR3 sequence in the samples, they observe that in almost every case the CDR3 sequences belong to a top clone. This is different from what was observed with colorectal cancer, where few of the high-scoring snippets appeared in the CDR3 sequences of the most abundance clones.

Table 3

COLORECTAL SAMPLES Sherwood *et al.*, 2013 [24]			
Patient # (Patient ID)	Tumor		Healthy
	MSI-status	Unique TCRBs	Unique TCRBs
1 (400464)	MSS	1,836	1,466
2 (400480)	MSS	2,432	1,773
3 (400488)	MSI-H	2,090	699
4 (400600)	MSS	862	984
5 (400712)	MSS	203	667
6 (400728)	MSS	41	1,110
7 (401144)	MSS	1,390	1,040
8 (401176)	MSS	723	883
9 (401248)	MSS	391	1,844
10 (401256)	MSS	1,711	1,068
11 (401264)	MSS	3,849	910
12 (401304)	MSS	1,659	1,612
13 (401320)	MSS	2,933	1,667
14 (401336)	MSI-H	988	1,228

Table 4

BREAST SAMPLES Beausang *et al.*, 2017 [25]			
Patient # (Patient ID)	Tumor		Healthy
	ER/PR/HER2	Unique TCRBs	Unique TCRBs
1 (BR01)	+ / + / -	50,667	18,848
2 (BR05)	+ / + / -	21,559	7,923
3 (BR07)	+ / + / -	22,345	12,334
4 (BR13)	+ / + / -	8,276	2,609
5 (BR14)	+ / + / -	34,203	5,577
6 (BR15)	+ / + / -	16,341	3,316
7 (BR16)	+ / + / -	8,237	22,483
8 (BR17)	+ / + / -	8,686	7,748
9 (BR18)	+ / + / -	5,324	812
10 (BR19)	+ / + / -	8,571	8,865
11 (BR20)	+ / + / -	15,956	13,611
12 (BR21)	+ / + / -	18,597	10,593
13 (BR22)	+ / + / -	51,097	22,774
14 (BR24)	- / - / -	45,953	10,903
15 (BR25)	- / - / +	16,004	4,276
16 (BR26)	+ / + / -	6,250	3,397

TABLE 5

Cancer Type	Computing the Relative Abundance of the Snippet from Relative Abundance of the Receptor	Log of the Relative Abundance, Normalized as a Feature	Initial Value for the Weigh Term on the Log of the Relative Abundance	Miscellaneous	Patient Holdout Cross-Validation
Colorectal Cancer SHERWOOD ET AL, 2013	Not Used	Un-normalized	$W_{21} = 0$		10/28 ≈ 36%
	Snippet	Un-normalized	$W_{21} = 0$		[redacted]
	Snippet	$\mu=0, \sigma=1$	$W_{21} \sim N(0, 1/21)$		18/28 ≈ 64%
	Receptor	Un-normalized	$W_{21} = 0$		21/28 ≈ 75%
	Receptor	Un-normalized	$W_{21} = 0$	Early Stopping	23/28 ≈ 82%
Breast Cancer BRAUSAN ET AL, 2017	Snippet	Un-normalized	$W_{21} = 0$		14/32 ≈ 44%
	Snippet	Un-normalized	$W_{21} = 0$	Early Stopping	23/23 ≈ 72%
	Snippet	Un-normalized	$W_{21} = 0$	Smaller Step Size	13/32 ≈ 41%
	Snippet	Un-normalized	$W_{21} = 0$	Smaller Step Size & Early Stopping	22/32 ≈ 69%
	Receptor	Un-normalized	$W_{21} = 0$		[redacted]
	Receptor	$\mu=0, \sigma=1$	$W_{21} = 0$		27/32 ≈ 84%
	Receptor	$\mu=0, \sigma=1$	$W_{21} = 0$		28/32 ≈ 87%

Several variations of the model are considered. (1st column) Cancer type. (2nd column) Two strategies are considered for computing the relative abundance of the snippet given the relative abundance of the receptor. The first method aggregates the relative abundances of each receptor containing the snippet. The second method identifies all receptors containing a snippet and uses the relative abundance of the most frequent receptor. (3rd column) The log of the relative abundance is a feature of the model and can either be normalized or not. (4th column) Different schemes for initializing the weight. (5th column) Other variations of the model that are considered, some of which are listed here. (5th column) The performance of each variation of the model.

* * * * * * * * * * * * *

All of the compositions and/or methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this disclosure have been described in terms of preferred embodiments, it will be apparent to those of skill in the art that variations may be applied to the compositions and/or methods and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the disclosure. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the disclosure as defined by the appended claims.

VII. References

The following references, to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference.

1. Luo, W., *et al.*, *Analysis of the interindividual conservation of T cell receptor alpha- and beta-chain variable regions gene in the peripheral blood of patients with systemic lupus erythematosus.* Clin Exp Immunol, 2008. **154**(3): p. 316-24.
2. Cameron, E.M., *et al.*, *Potential of a unique antibody gene signature to predict conversion to clinically definite multiple sclerosis.* J Neuroimmunol, 2009. **213**(1-2): p. 123-30.
3. Marrero, I., D.E. Hamm, and J.D. Davies, *High-throughput sequencing of islet-infiltrating memory CD4+ T cells reveals a similar pattern of TCR Vbeta usage in prediabetic and diabetic NOD mice.* PLoS One, 2013. **8**(10): p. e76546.
4. Iglesia, M.D., *et al.*, *Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer.* Clin Cancer Res, 2014. **20**(14): p. 3818-29.
5. Jia, Q., *et al.*, *Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer.* Oncoimmunology, 2015. **4**(4): p. e1001230.
6. Postow, M.A., *et al.*, *Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma.* J Immunother Cancer, 2015. **3**: p. 23.
7. Rounds, W.H., *et al.*, *MSPrecise: A molecular diagnostic test for multiple sclerosis using next generation sequencing.* Gene, 2015. **572**(2): p. 191-7.
8. Polman, C.H., *et al.*, *Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria.* Ann Neurol, 2011. **69**(2): p. 292-302.
9. Frohman, E.M., *et al.*, *Most patients with multiple sclerosis or a clinically isolated demyelinating syndrome should be treated at the time of diagnosis.* Arch Neurol, 2006. **63**(4): p. 614-9.
10. Atchley, W.R., *et al.*, *Solving the protein sequence metric problem.* Proc Natl Acad Sci U S A, 2005. **102**(18): p. 6395-400.

11. Kingma, D. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

12. Robinson, W.H., *Sequencing the functional antibody repertoire--diagnostic and therapeutic discovery.* Nat Rev Rheumatol, 2015. **11**(3): p. 171-82.

13. Milo, R. and A. Miller, *Revised diagnostic criteria of multiple sclerosis.* Autoimmun Rev, 2014. **13**(4-5): p. 518-24.

14. Kim, S.S., *et al.*, *Limited utility of current MRI criteria for distinguishing multiple sclerosis from common mimickers: primary and secondary CNS vasculitis, lupus and Sjogren's syndrome.* Mult Scler, 2014. **20**(1): p. 57-63.

15. Owens, G.P., *et al.*, *VH4 gene segments dominate the intrathecal humoral immune response in multiple sclerosis.* J Immunol, 2007. **179**(9): p. 6343-51.

16. Bennett, J.L., *et al.*, *CSF IgG heavy-chain bias in patients at the time of a clinically isolated syndrome.* J Neuroimmunol, 2008. **199**(1-2): p. 126-32.

17. Yaari, G. and S.H. Kleinstein, *Practical guidelines for B-cell receptor repertoire sequencing analysis.* Genome Med, 2015. **7**: p. 121.

18. Vander Heiden, J.A., *et al.*, *pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires.* Bioinformatics, 2014. **30**(13): p. 1930-2.

19. Ye, J., *et al.*, *IgBLAST: an immunoglobulin variable domain sequence analysis tool.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W34-40.

20. Quail, M.A., *et al.*, *SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing.* BMC Genomics, 2014. **15**: p. 110.

21. Seitz, V., *et al.*, *A new method to prevent carry-over contaminations in two-step PCR NGS library preparations.* Nucleic Acids Res, 2015. **43**(20): p. e135.

22. Abadi, M., *et al.*, *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, 2016.

23. Atchley, William R., *et al.*, *Solving the protein sequence metric problem.* Proceedings of the National Academy of Sciences of the United States of America 102.18 (2005): 6395-6400.

24. Sherwood, Anna M., *et al.*, *Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue.* Cancer Immunology, Immunotherapy 62.9 (2013): 1453-1461.

25. Beausang, John F., *et al.*, *T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the T cell repertoire.* Proceedings of the National Academy of Sciences (2017): 201713863.

What is Claimed:

1. A method of identifying a disease biomarker from adaptive immune receptor sequences comprising:

(a) obtaining the sequence encoding one or more adaptive immune receptors from a plurality of immune cells obtained from (i) a plurality of subjects having a given disease and (ii) a plurality of control subjects;

(b) assessing the following biochemical properties for each amino acid lying in a plurality of arbitrarily defined regions or subregions within said one or more adaptive immune receptor sequences:

(i) polarity;
(ii) secondary structure;
(iii) molecular volume;
(iv) codon diversity, and
(v) electrostatic charge,

(c) selecting one or more regions or subregions within said one or more adaptive immune receptor sequences;

(d) scoring each region or subregion based on the biochemical properties using a parameterized detector function;

(e) aggregating the scores from a patient's plurality of said regions or subregions to predict a patient diagnosis; and

(f) adjusting the parameters of the scoring function to yield the correct diagnosis for each patient in the example data, thereby identifying an adaptive immune receptor-related disease biomarker.

2. The method of claim 1, further comprising assessing the regions or subregions in step (b) in combination with the logarithm of the relative abundance (also known as frequency count), where the relative abundance is either:

(i) the relative abundance of the most abundant receptor containing the subregions or regions, or

(ii) the relative abundance of each receptor containing the subregions or regions, or

(iii) the relative abundance of the subregions or regions, which can be calculated by summing the abundances of all the CDR3 sequences containing the subregions or regions and dividing by the total count of all subregions or regions.

3. The method of claim 1, wherein step (a) comprises amplification of said sequence.

4. The method of claims 1-3, wherein step (a) comprises any high-throughput sequencing platform, including but not limited to 454 or Illumina sequencers.

5. The method of claims 1-4, wherein the disease is a human or animal disease, syndrome, or disorder in which lymphocytes play a role.

6. The method of claims 1-5, wherein only heavy chain, beta chain, or delta chain CDR3 is analyzed.

7. The method of claims 1-6, wherein CDR coding sequences are obtained from a VH4 immunoglobulin.

8. The method of claims 1-5, wherein one or more light, alpha, or gamma chain CDRs are analyzed.

9. The method of claims 1-5, wherein one or more CDRs are from B cell or T cell receptors.

10. The method of claim 1, wherein the one or more subregions each consist of between 5 and 10 codons.

11. The method of claim 10, wherein the one or more subregions each consist of 6 codons.

12. The method of claim 1, wherein the detector function is a logistic regression function.

13. The method of claim 1, wherein the detector function is not a logistic regression function.

14. The method of claim 1, wherein the scores from a patient's plurality of said regions or subregions are aggregated together, such as by taking the highest score among the plurality of scores.

15. The method of claim 1, wherein the biomarker is used to diagnose and/or treat a patient.

16. A method of identifying a subject as having or at risk of developing multiple sclerosis comprising:

(a) obtaining the sequence encoding one or more heavy chain CDR3s from a plurality of B cells obtained from a subject;

(b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

DFNWFD (SEQ ID NO: 1)
IMKWFD (SEQ ID NO: 2)
DGSWAE (SEQ ID NO: 3)
DVWKAP (SEQ ID NO: 4)
DFWNEV (SEQ ID NO: 5)
RQRYLD (SEQ ID NO: 6)
DKNWLD (SEQ ID NO: 7)
NCHPFD (SEQ ID NO: 8)
HLNWFD (SEQ ID NO: 9)
QLFWFD (SEQ ID NO: 10)
EPQDAF (SEQ ID NO: 11)
LYHYDS (SEQ ID NO: 12)
DYWYLD (SEQ ID NO: 13)
DYWYFD (SEQ ID NO: 14)
WYLDLW (SEQ ID NO: 15)
WYFDLW (SEQ ID NO: 16)
EEQWLA (SEQ ID NO: 17)

KQQQRF (SEQ ID NO: 18)
DYSYFD (SEQ ID NO: 19)
SEWYID (SEQ ID NO: 20)
QTQSIV (SEQ ID NO: 21)
DCHYFD (SEQ ID NO: 22)
DWEWLL (SEQ ID NO: 23)
DVEWLL (SEQ ID NO: 24)
WEWLLF (SEQ ID NO: 25)
EWLFFD (SEQ ID NO: 26)
EWLLFD (SEQ ID NO: 27)
DLHHHY (SEQ ID NO: 28)
DLHCHY (SEQ ID NO: 29)
HYHYVM (SEQ ID NO: 30)
DLHYHY (SEQ ID NO: 31)
ELHYHY (SEQ ID NO: 32)
HHHYGM (SEQ ID NO: 33)
HPHDAF (SEQ ID NO: 34)
FCHPHD (SEQ ID NO: 35)
DAFDLW (SEQ ID NO: 36)
KFWDLL (SEQ ID NO: 37)
AIRHSD (SEQ ID NO: 38)
AVRHSD (SEQ ID NO: 39)
HLLLLH (SEQ ID NO: 40)
REHMAV (SEQ ID NO: 41)
WYLDLW (SEQ ID NO: 42)
WYFDLW (SEQ ID NO: 43)
EYFQHW (SEQ ID NO: 44)
HTNFDD (SEQ ID NO: 45)
WYFYLW (SEQ ID NO: 46)
HWRHCS (SEQ ID NO: 47)
HVRHCS (SEQ ID NO: 48)
SFHFDS (SEQ ID NO: 49)
ARHWRH (SEQ ID NO: 50)
HGRHCS (SEQ ID NO: 51)

HYYMDV (SEQ ID NO: 52); and

(c) identifying said subject as having or at risk of developing multiple sclerosis when one of more of said sequences is identified.

17. The method of claim 16, wherein CDR coding sequences are obtained from a VH4 immunoglobulin.

18. The method of claim 16, wherein step (a) comprises amplification of said sequence.

19. The method of claim 16, wherein step (a) comprises high-throughput sequencing platform including but not limited to 454 or Illumina sequencers.

20. The method of claim 16, further comprising providing to said subject a therapeutic treatment for multiple sclerosis.

21. The method of claim 16, further comprising providing to said subject a prophylactic treatment for multiple sclerosis.

22 The method of claim 16, wherein said subject is suspected of having an autoimmune disease.

23. The method of claim 16, wherein said subject is suspected of having multiple sclerosis.

24. The method of claim 16, wherein said subject has previously been diagnosed as having multiple sclerosis.

25. The method of claims 20 or 21, further comprising performing steps (a)-(c) a second time after said treatment to assess a change in the B cell repertoire.

26. A method of identifying a subject as having or at risk of developing colorectal cancer comprising:

(a) obtaining the sequence encoding one or more beta chain CDR3s from a plurality of T cells obtained from a subject;

(b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

MGRM (SEQ ID NO: 53)
IRQM (SEQ ID NO: 54)
ENRI (SEQ ID NO: 55)
GRHM (SEQ ID NO: 56)
IRDM (SEQ ID NO: 57)
RGKM (SEQ ID NO: 58)
IGRM (SEQ ID NO: 59)
INKI (SEQ ID NO: 60)
HREF (SEQ ID NO: 61)
RRTM (SEQ ID NO: 62)
ERRM (SEQ ID NO: 63)
ERRM (SEQ ID NO: 64)
HNRM (SEQ ID NO: 65)
IRKE (SEQ ID NO: 66)
HGRM (SEQ ID NO: 67)
YREF (SEQ ID NO: 68)
WKDY (SEQ ID NO: 69)
MYRE (SEQ ID NO: 70)
YREV (SEQ ID NO: 71)
ERFY (SEQ ID NO: 72)
RERF (SEQ ID NO: 73)
MRGM (SEQ ID NO: 74)
ERSI (SEQ ID NO: 75)
IRQF (SEQ ID NO: 76)
RRHI (SEQ ID NO: 77); and

(c) identifying said subject as having or at risk of developing colorectal cancer when one of more of said sequences is identified.

27. A method of identifying a subject as having or at risk of developing breast cancer comprising:

(a) obtaining the sequence encoding one or more beta chain CDR3s from a plurality of T cells obtained from a subject;

(b) identifying one or more of sequences in said one or more CDR3s selected from the group consisting of:

LSRG (SEQ ID NO: 78)
LSRS (SEQ ID NO: 79)
RSNQ (SEQ ID NO: 80)
LSYE (SEQ ID NO: 81)
ASYN (SEQ ID NO: 82)
AGNQ (SEQ ID NO: 83)
GSYN (SEQ ID NO: 84)
ASNQ (SEQ ID NO: 85)
LCNN (SEQ ID NO: 86)
ASYE (SEQ ID NO: 87)
SSYN (SEQ ID NO: 88)
LPRD (SEQ ID NO: 89)
SSYN (SEQ ID NO: 90)
LDGQ (SEQ ID NO: 91)
PSNQ (SEQ ID NO: 92)
ASNE (SEQ ID NO: 93)
AYNQ (SEQ ID NO: 94)
AAYN (SEQ ID NO: 95)
SSPH (SEQ ID NO: 96)
DSNQ (SEQ ID NO: 97)
SSNN (SEQ ID NO: 98)
SSYE (SEQ ID NO: 99)
ASNQ (SEQ ID NO: 100)
SSYN (SEQ ID NO: 101)
ASRD (SEQ ID NO: 102)
SSKD (SEQ ID NO: 103); and

(c) identifying said subject as having or at risk of developing breast cancer when one of more of said sequences is identified.

28. The method of claim 26 or 27, wherein step (a) comprises amplification of said sequence.

29. The method of claim 26 or 27, wherein step (a) comprises any high-throughput sequencing platform including but not limited to 454 or Illumina sequencers.

30. The method of claim 26 or 27, further comprising providing to said subject a therapeutic treatment for cancer.

31. The method of claim 26 or 27, further comprising providing to said subject a prophylactic treatment for cancer.

32. The method of claim 26 or 27, wherein said subject is suspected of having cancer.

33. The method of claim 31, wherein said subject is suspected of having colorectal or breast cancer.

34. The method of claim 26, wherein said subject has previously been diagnosed as having cancer.

35. The method of claim 33, wherein said subject has previously been diagnosed as having colorectal or breast cancer.

36. The method of claims 30 or 31, further comprising performing steps (a)-(c) a second time after said treatment to assess a change in the T cell repertoire.

Collect B-cells from Patient

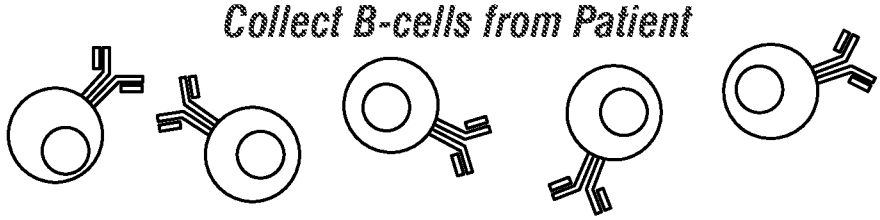


FIG. 1A

Next Generation Sequencing

Antibody Sequences

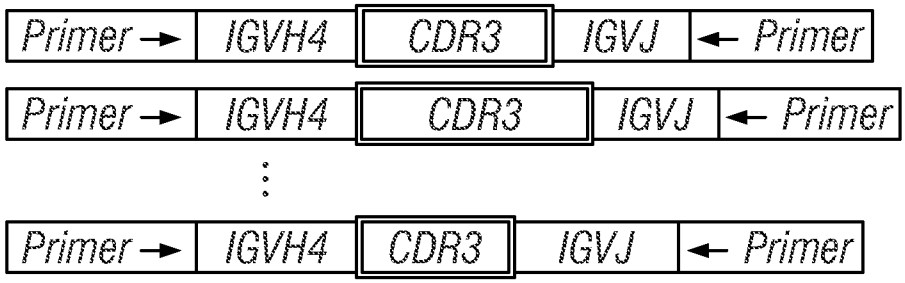
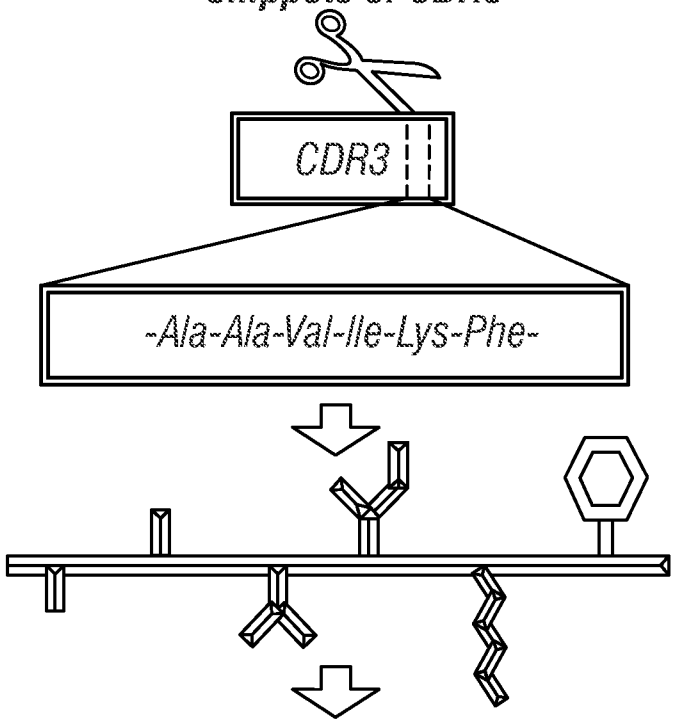


FIG. 1B

Snippets of CDR3



Biochemical Features
(ATCHLEY FACTORS)

$$\begin{pmatrix} -0.6 \\ -1.3 \\ -0.7 \\ 1.6 \\ -0.1 \end{pmatrix} \begin{pmatrix} -0.6 \\ -1.3 \\ -0.7 \\ 1.6 \\ -0.1 \end{pmatrix} \begin{pmatrix} -1.3 \\ -0.3 \\ -0.5 \\ 1.2 \\ -1.3 \end{pmatrix} \begin{pmatrix} -1.2 \\ -0.5 \\ 2.1 \\ 0.4 \\ 0.8 \end{pmatrix} \begin{pmatrix} -1.8 \\ -0.6 \\ 0.5 \\ -0.3 \\ 1.6 \end{pmatrix} \begin{pmatrix} -1.0 \\ -0.6 \\ 1.9 \\ -0.4 \\ 0.4 \end{pmatrix}$$

FIG. 1C

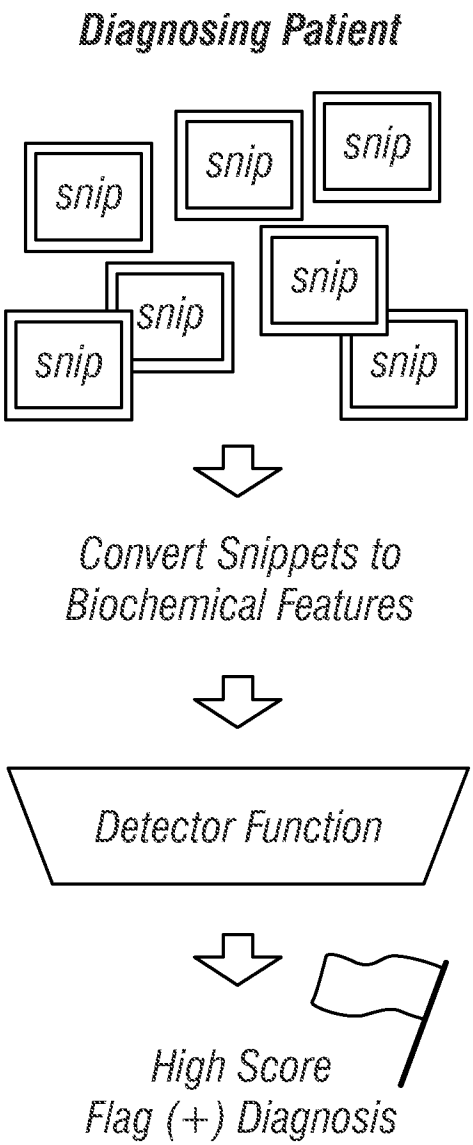

Diagnosing Patient
snip
snip
snip
snip
snip
snip
snip
snip
Convert Snippets to
Biochemical Features
Detector Function
High Score
Flag (+) Diagnosis

FIG. 1D

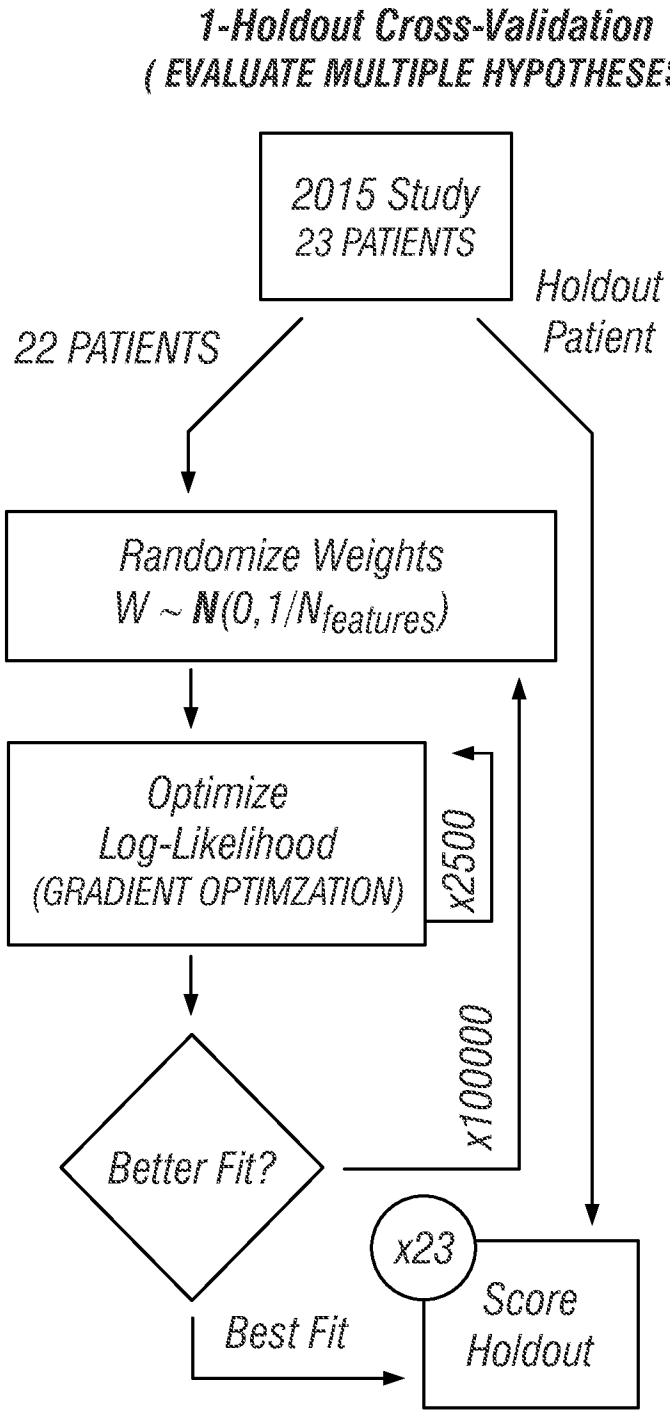

1-Holdout Cross-Validation
(EVALUATE MULTIPLE HYPOTHESES)
2015 Study
23 PATIENTS
22 PATIENTS
Holdout
Patient
Randomize Weights
W ~ N(0,1/N features)
Optimize
Log-Likelihood
(GRADIENT OPTIMZATION)
x2500
x100000
Better Fit?
x23
Score
Holdout
Best Fit

FIG. 2A

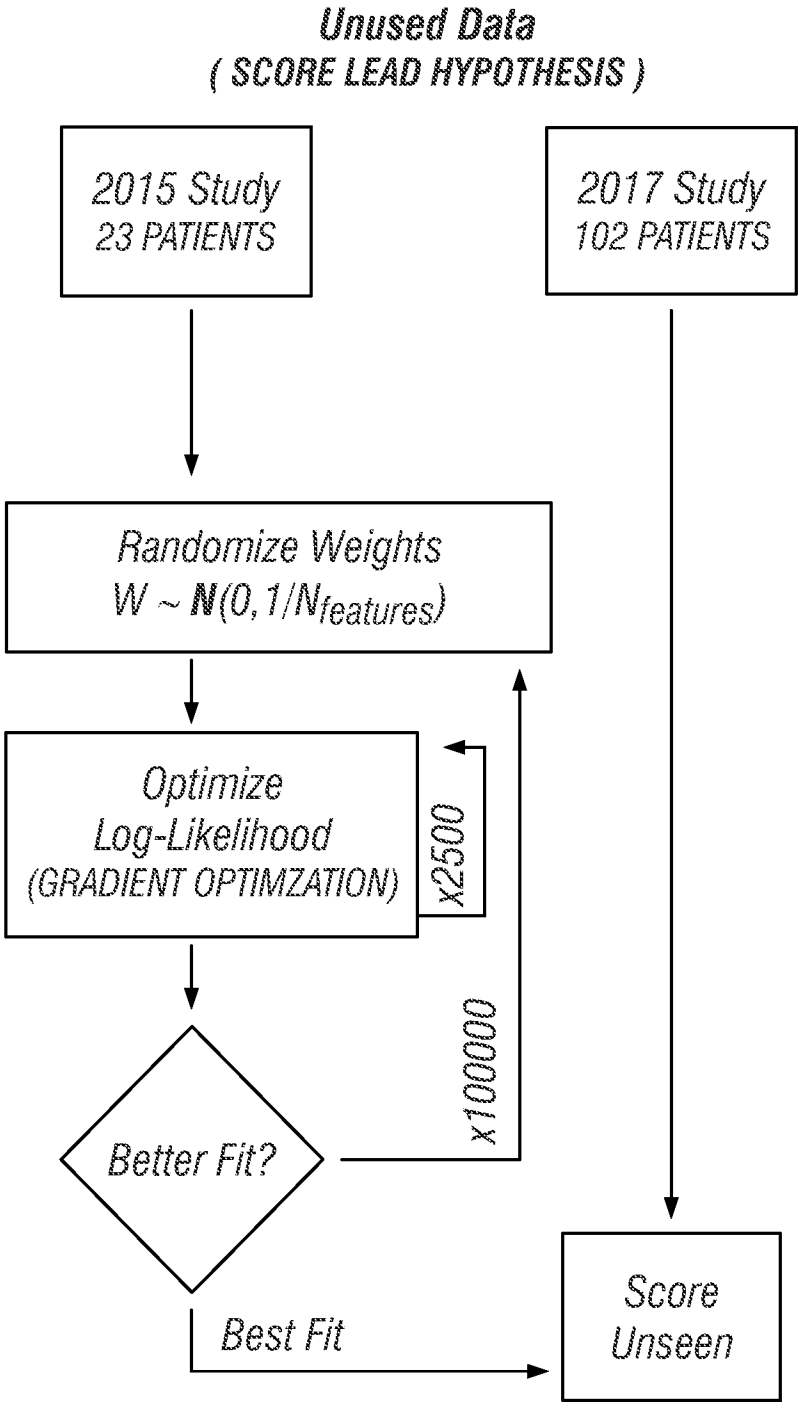

Unused Data
(SCORE LEAD HYPOTHESIS)
2015 Study
23 PATIENTS
2017 Study
102 PATIENTS
Randomize Weights
W ~ N(0,1/N features)
Optimize
Log-Likelihood
(GRADIENT OPTIMZATION)
x2500
Better Fit?
x100000
Best Fit
Score
Unseen

FIG. 2B

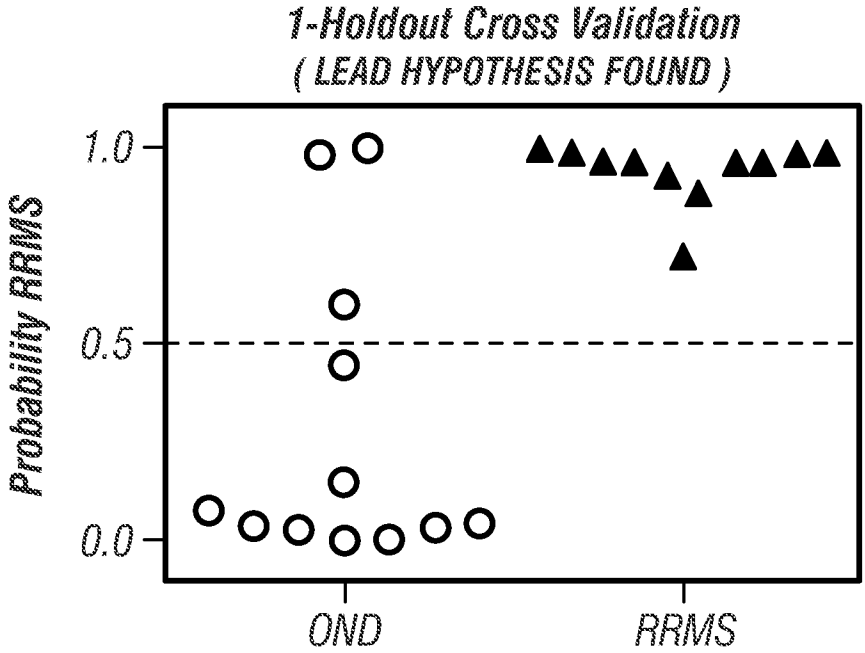

1-Holdout Cross Validation
(LEAD HYPOTHESIS FOUND)
Probability RRMS
1.0
0.5
0.0
OND
RRMS

FIG. 3A

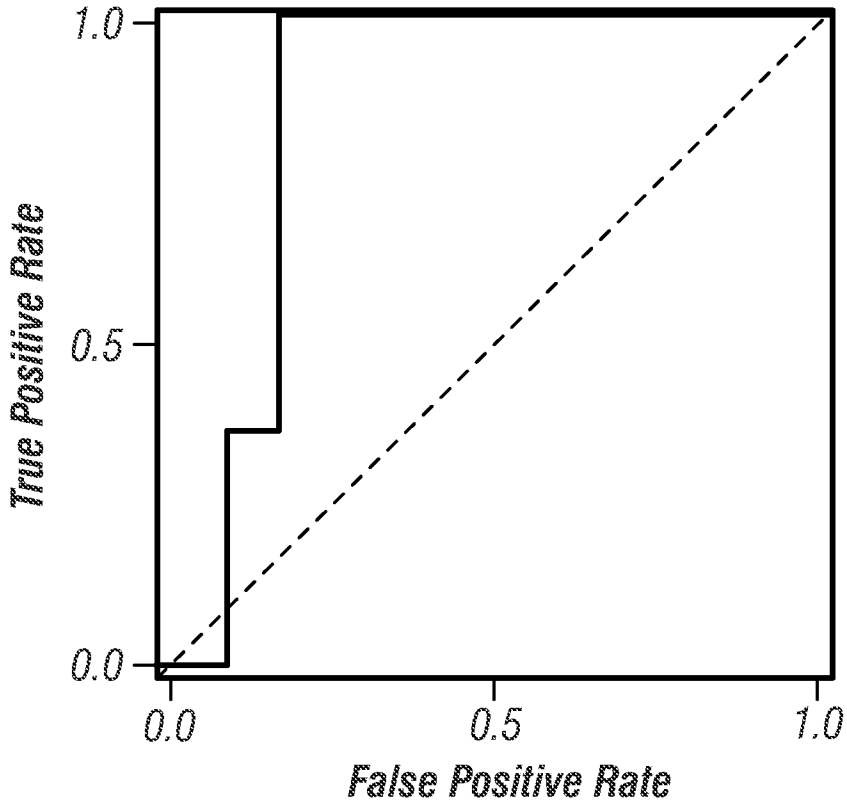

True Positive Rate
1.0
0.5
0.0
0.0
0.5
1.0
False Positive Rate

FIG. 3B

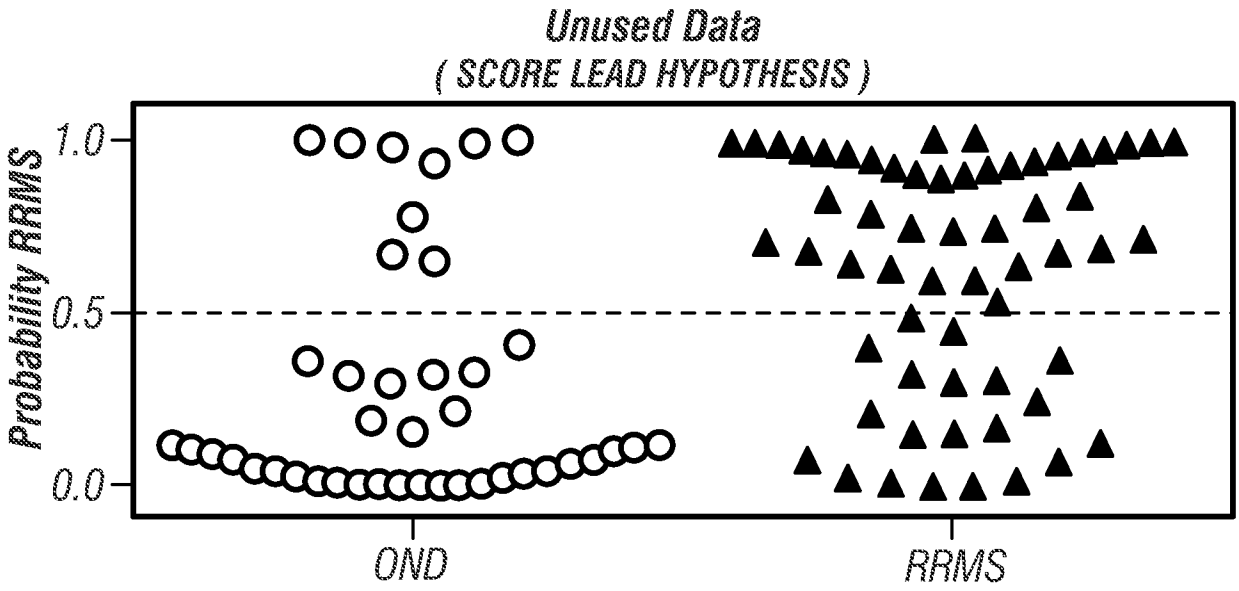

Unused Data
(SCORE LEAD HYPOTHESIS)
Probability RRMS
1.0
0.5
0.0
OND
RRMS

FIG. 3C

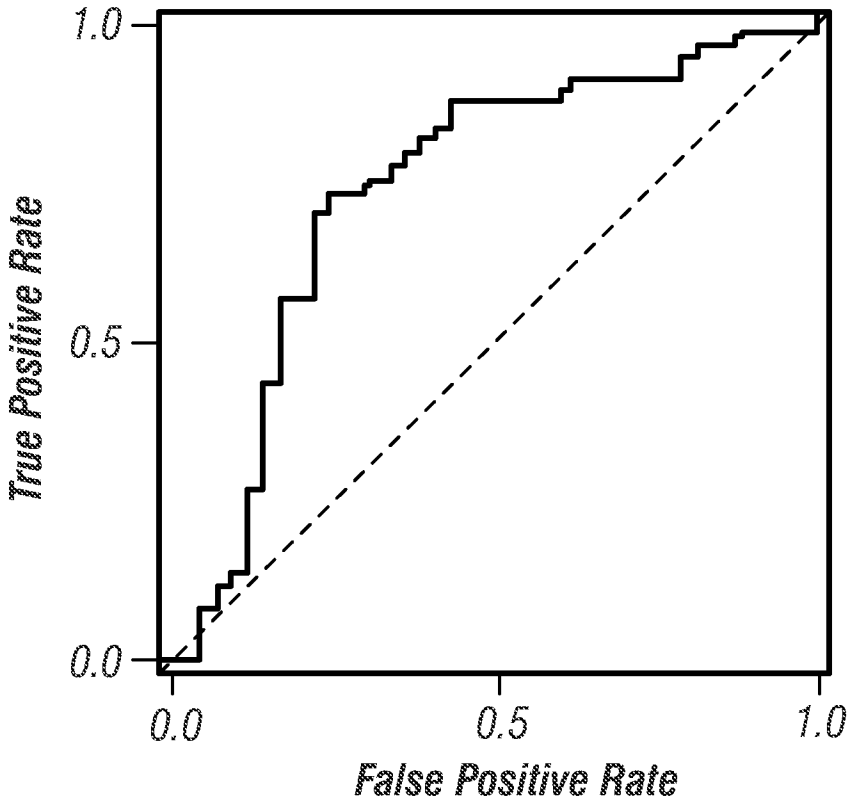

True Positive Rate
1.0
0.5
0.0
0.0
0.5
1.0
False Positive Rate

FIG. 3D

Model Weights
i.
Hydrophilicity
ii.
Bends, Coils, Turns
iii.
Small Size
iv.
Higher Usage
v.
Positively Charged
Detector Function Weights
Hydrophobicity
α-helical
Large Size
Heat Capacity, Refractivity
Negatively Charged

FIG. 4

Top Atchley Factor Correlations
i.
Hydrophilicity
ii.
Bends, Coils, Turns
iii.
Small Size
iv.
Higher Usage in All Proteins
v.
Positively Charged
Hydrophobicity
α-helical
Large Size
Heat Capacity, Refractivity
Negatively Charged

FIG. 5

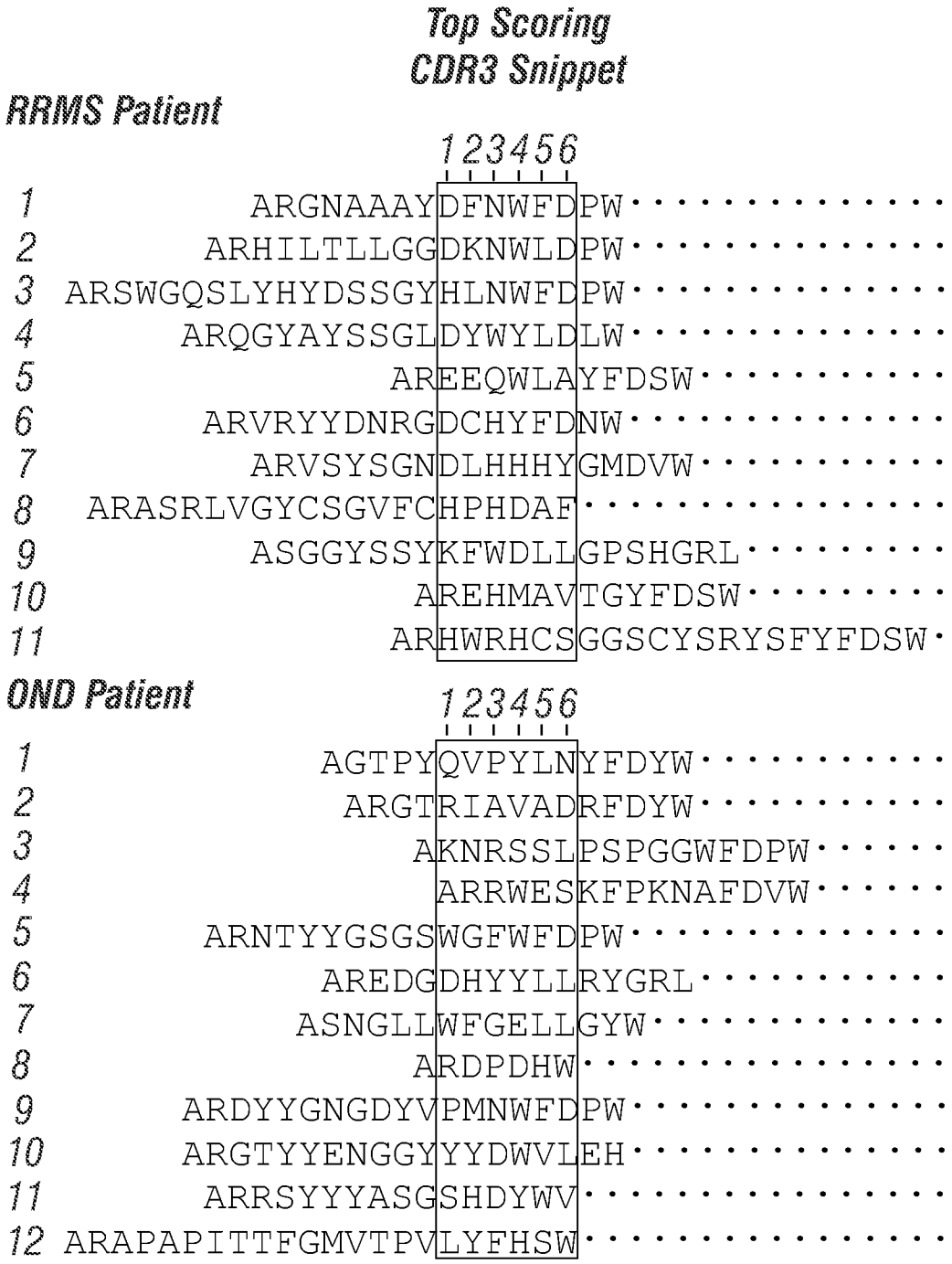

Top Scoring
CDR3 Snippet
RRMS Patient
1 2 3 4 5 6
1 ARGNAAAYDFNWFDPW
2 ARHILTLLGGDKNWLDPW
3 ARSWGQSLYHYDSSGYHLNWFDPW
4 ARQGYAYSSGLDYWYLDLW
5 AREEQWLAYFDSW
6 ARVRYYDNRGDCHYFDNW
7 ARVSYSGNDLHHHYGMDVW
8 ARASRLVGYCSGVFCHPHDAF
9 ASGGYSSYKFWDLLGPSHGRL
10 AREHMAVTGYFDSW
11 ARHWRHCSGGSCYSRYSFYFDSW
OND Patient
1 2 3 4 5 6
1 AGTPYQVPYLNYFDYW
2 ARGTRIAVADRFDYW
3 AKNRSSLPSPGGWFDPW
4 ARRWESKFPKNAFDVW
5 ARNTYYGSGSWGFWFDPW
6 AREDGDHYYLLRYGRL
7 ASNGLLWFGELLGYW
8 ARDPDHW
9 ARDYYGNGDYVPMNWFDPW
10 ARGTYYENGGYYYDWVLEH
11 ARRSYYYASGSHDYWV
12 ARAPAPITTFGMVTPVLYFHSW

FIG. 5
(Cont'd)

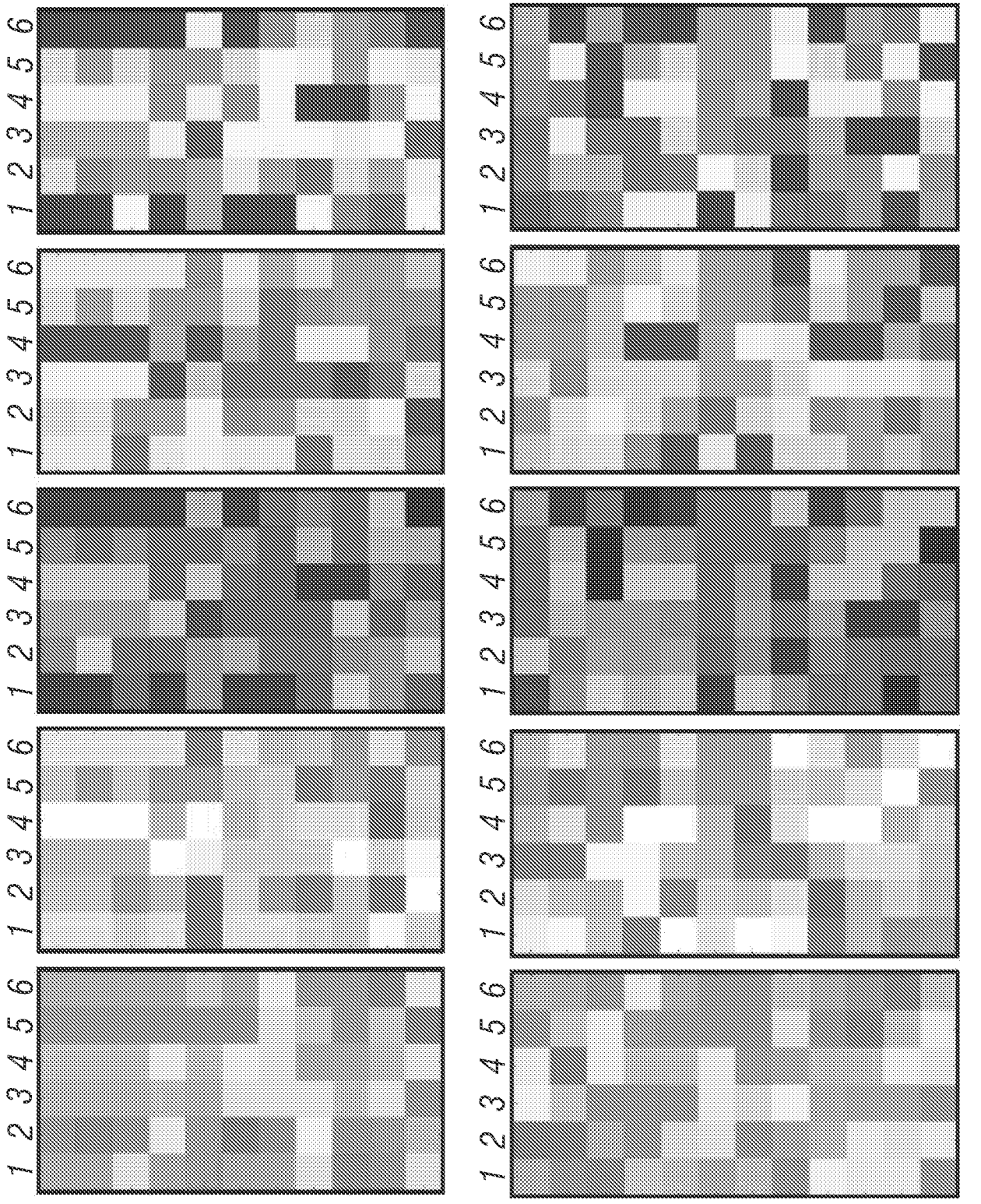

1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6
1 2 3 4 5 6

FIG. 5 (Cont'd)

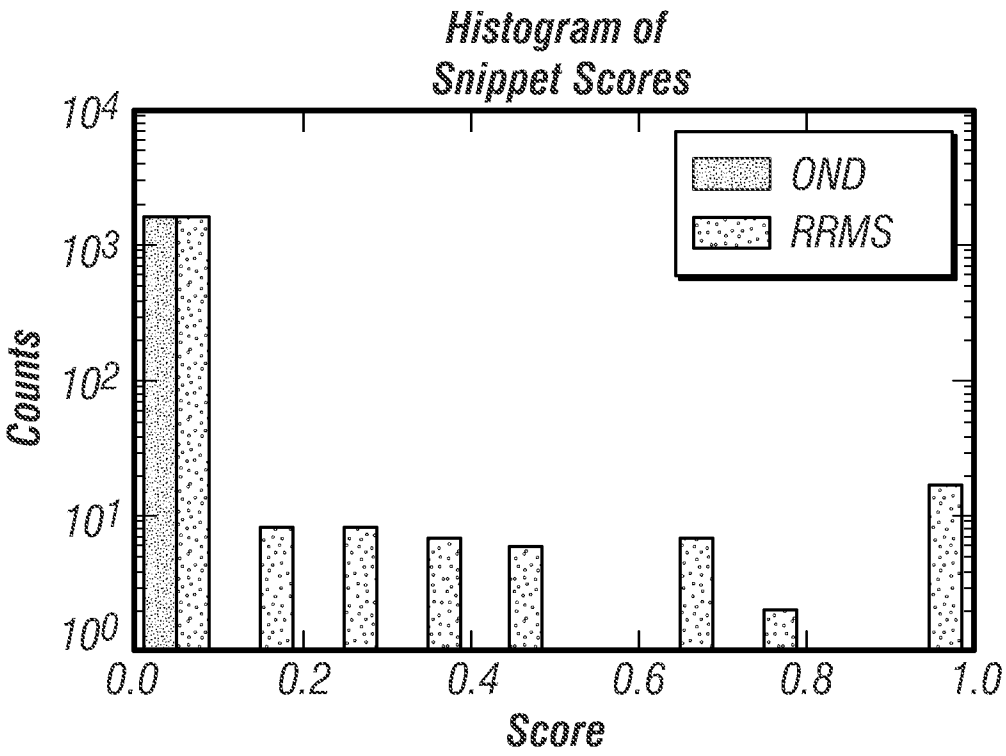


FIG. 6

Snippet	H	W	R	H	C	S	
DNA Encodings	CAT	TGG	CGT	CAT	TGT	TCT	
	CAC		CGC	CAC	TGC	TCC	
			CGA			TCA	
			CGG			TCG	
			AGA			AGT	
			AGG			AGC	
	x 2	x 1	x 6	x 2	x 2	x 6	= 288

FIG. 7

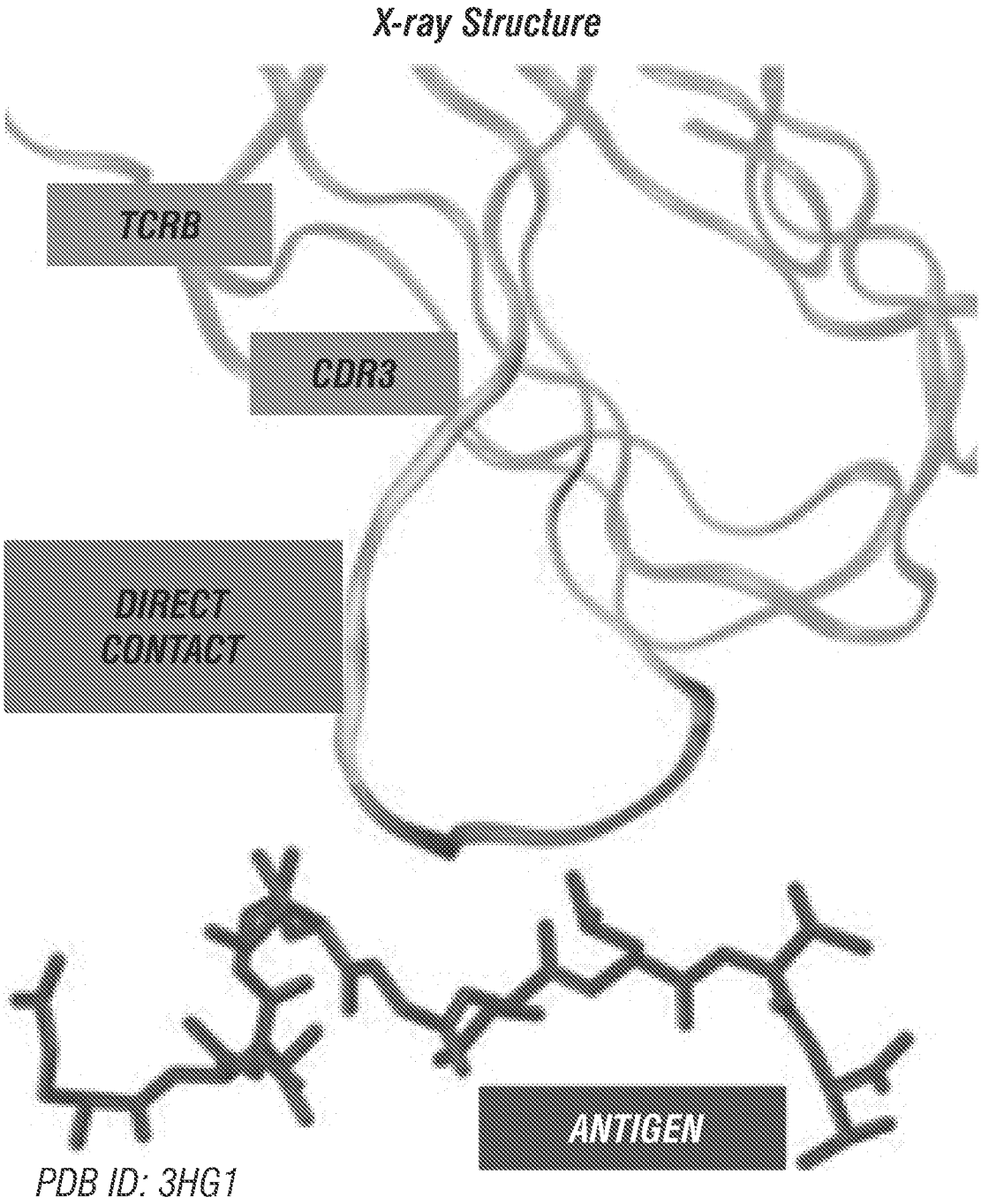

X-ray Structure
TCRB
CDR3
DIRECT CONTACT
ANTIGEN
PDB ID: 3HG1

FIG. 8A

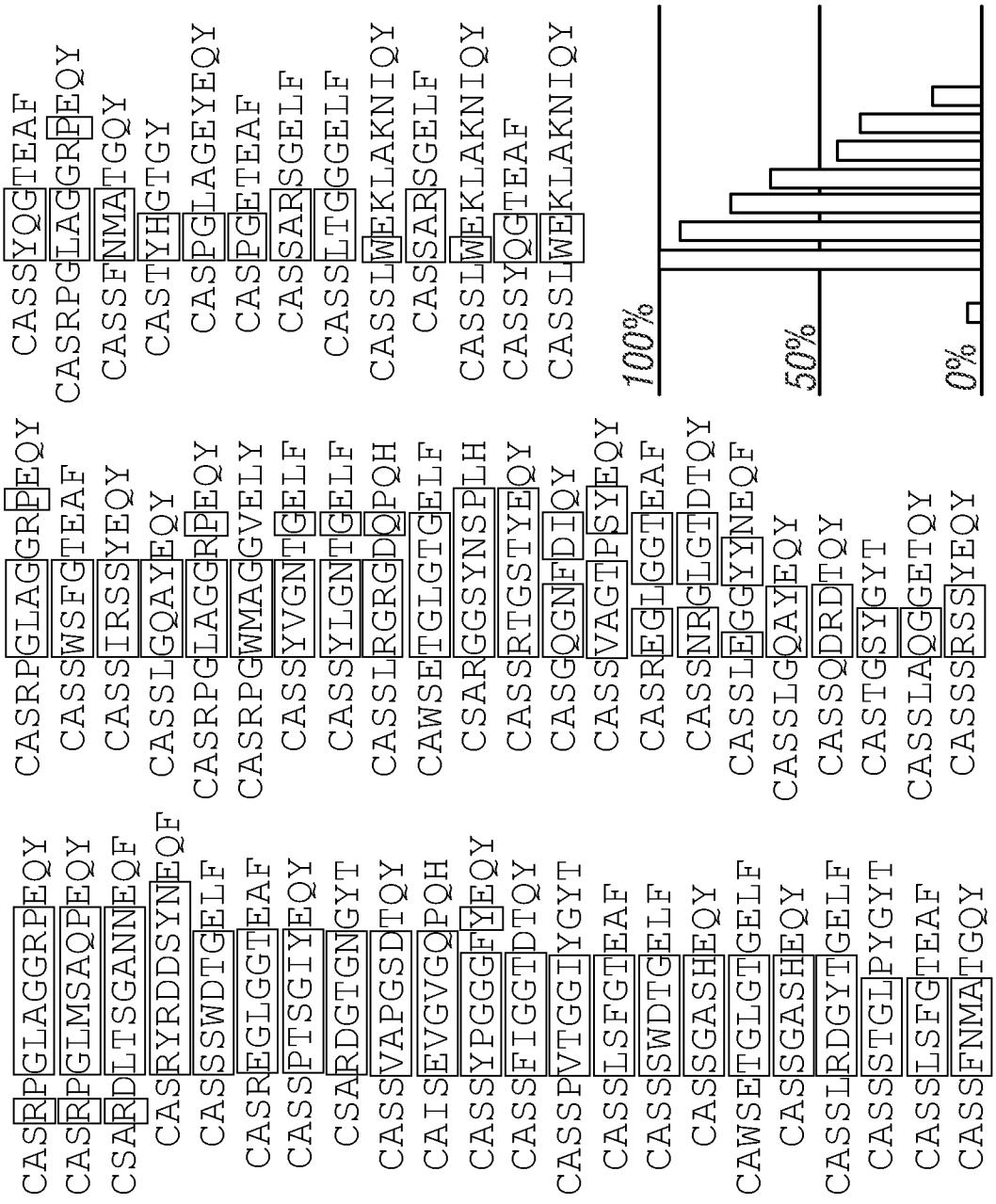

CDR3 from X-ray Structures Aligned by Contact Sites
CASRPGLAGGRPEQY
CASRPGLMSAQPEQY
CSARDLTSGANNEQF
CASRYRDDSYNEQF
CASSSWDTGELF
CASREGLGGTEAF
CASSPTSGIYEQY
CSARDGTGNGYT
CASSVAPGSDTQY
CAISEVGVGQPQH
CASSYPGGGFYEQY
CASSFIGGTDTQY
CASPVTGGIYGYT
CASSLSFGTEAF
CASSWDTGELF
CASSGASHEQY
CAWSETGLGTGELF
CASSGASHEQY
CASSLRDGYTGELF
CASSTGLPYGYT
CASSLSFGTEAF
CASSFNMATGQY
CASRPGLAGGRPEQY
CASSWSFGTEAF
CASSIRSSYEQY
CASSLGQAYEQY
CASRPGLAGGRPEQY
CASRPGWMAGGVELY
CASSYVGNTGELF
CASSYLGNTGELF
CASSLRGRGDQPQH
CAWSETGLGTGELF
CSARGGSYNSPLH
CASSRTGSTYEQY
CASGQGNFDIQY
CASSVAGTPSYEQY
CASREGLGGTEAF
CASSNRGLGTDTQY
CASSLEGGYYNEQF
CASSLGQAYEQY
CASSQDRDTQY
CASTGSYGYT
CASSLAQGGETQY
CASSRSSYEQY
CASSYQGTEAF
CASRPGLAGGRPEQY
CASSFNMATGQY
CASTYHGTGY
CASPGLAGEYEQY
CASPGETEAF
CASSARSGELF
CASSLTGGGELF
CASSLWEKLAKNIQY
CASSARSGELF
CASSLWEKLAKNIQY
CASSYQGTEAF
CASSLWEKLAKNIQY
100%
50%
0%

FIG. 8B

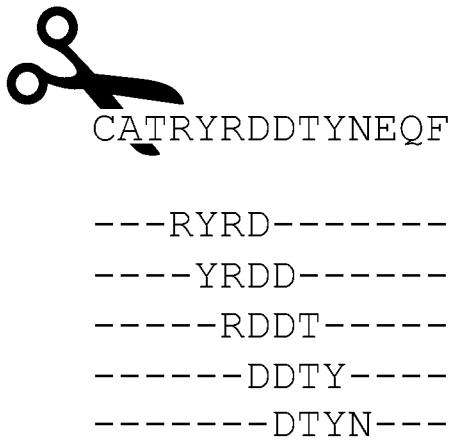

CDR3 Cut Into Snippets
CATRYRDDTYNEQF
---RYRD-------
----YRDD------
-----RDDT-----
------DDTY----
-------DTYN---

FIG. 8C

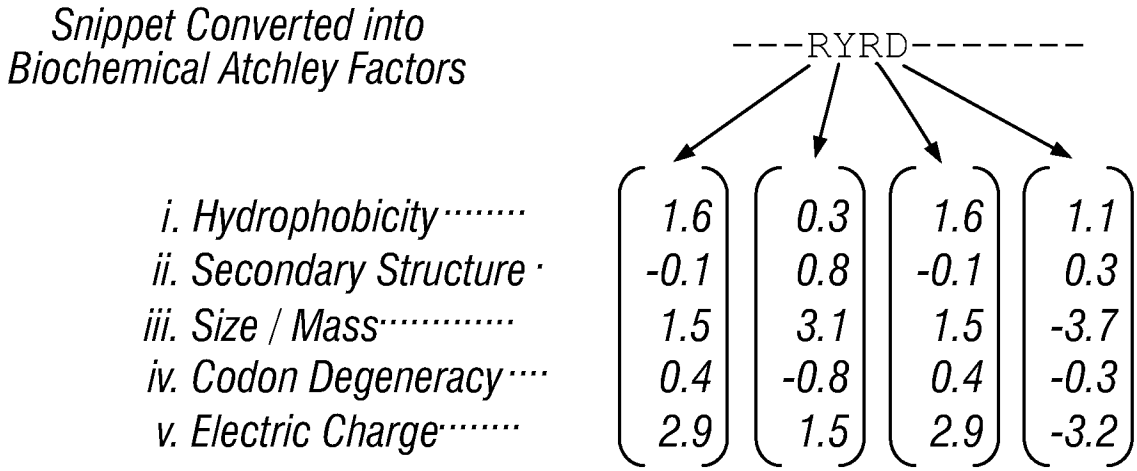

Snippet Converted into Biochemical Atchley Factors
---RYRD-------
i. Hydrophobicity
ii. Secondary Structure
iii. Size / Mass
iv. Codon Degeneracy
v. Electric Charge
1.6 0.3 1.6 1.1
-0.1 0.8 -0.1 0.3
1.5 3.1 1.5 -3.7
0.4 -0.8 0.4 -0.3
2.9 1.5 2.9 -3.2

FIG. 8D

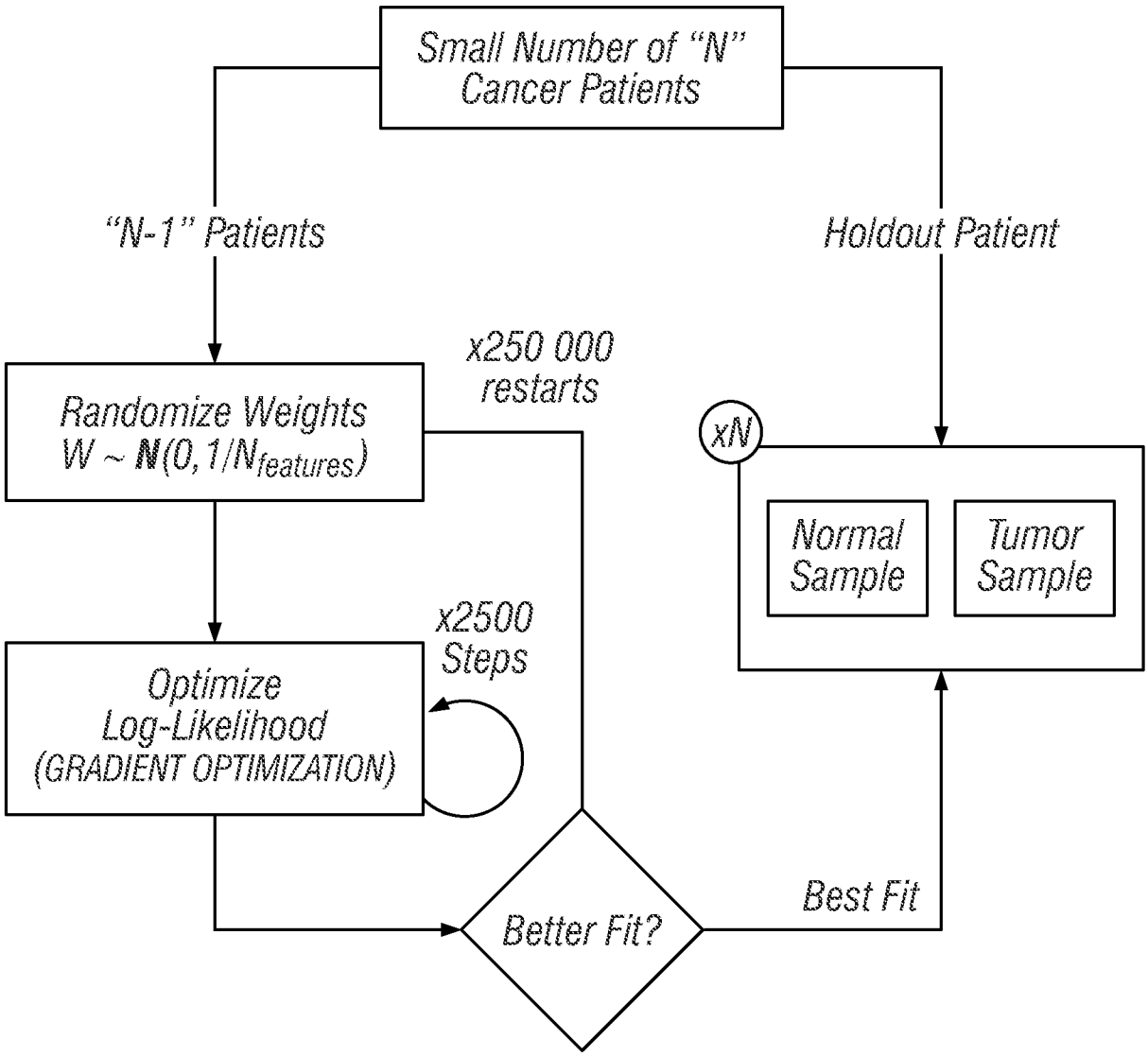

Small Number of "N" Cancer Patients
"N-1" Patients
Holdout Patient
x250 000 restarts
Randomize Weights
W ~ N(0,1/N features)
xN
Normal Sample
Tumor Sample
x2500 Steps
Optimize Log-Likelihood (GRADIENT OPTIMIZATION)
Better Fit?
Best Fit

FIG. 9

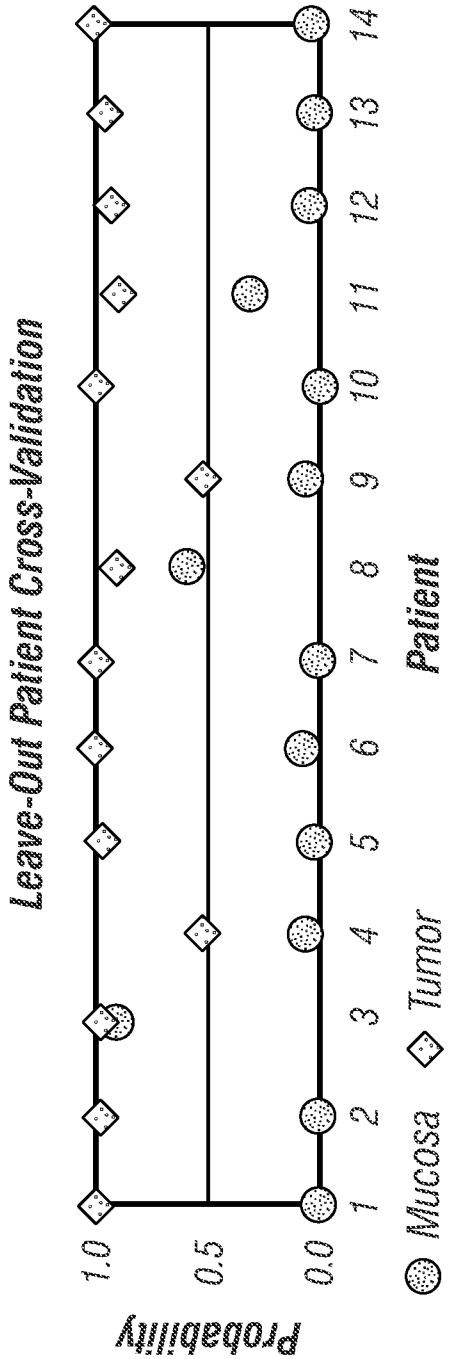


FIG. 10A

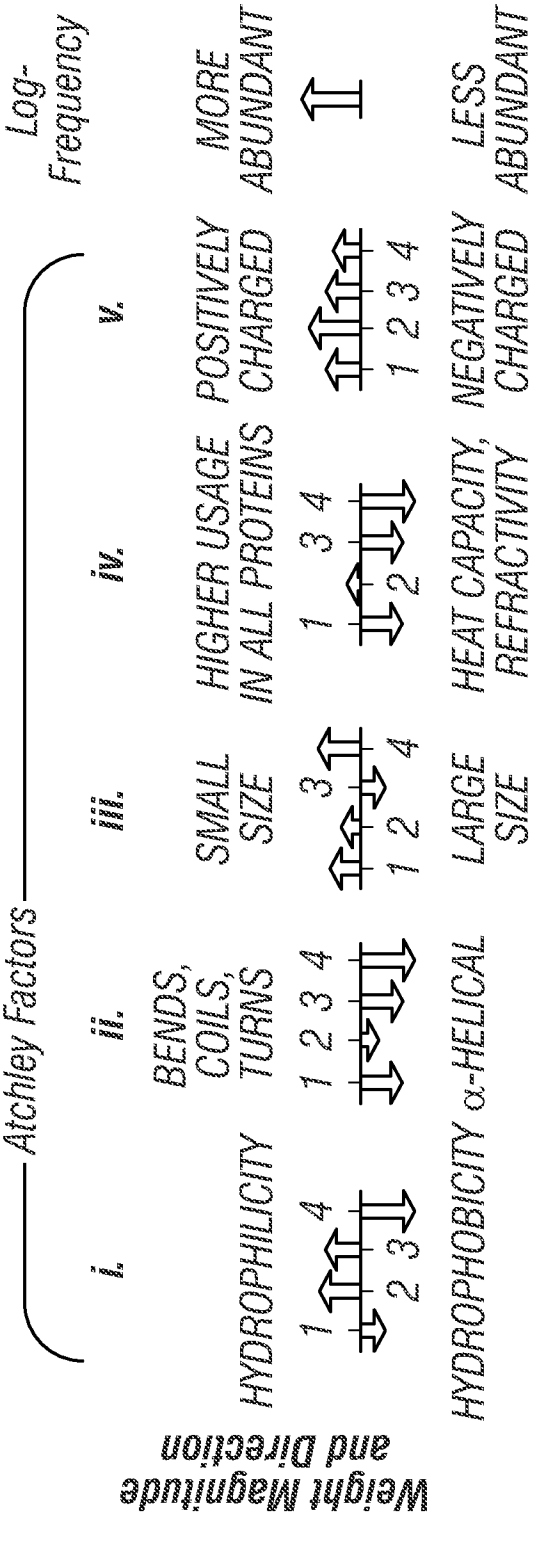


FIG. 10B

High Scoring Snippets

Patient	Snippet in CDR3	Rank
	1234	
1	CASS----BRRM--------NTEAFF	98
2	CASSL---BRRM--------NTEAFF	297
3	CASS----IRKE-------VLTEAFF	5
4	CASSFD--HGRM--------NTEAFF	91
	CASS----YREF------VRSGELFF	318
5	CASSF---WKDY--------QETQYF	1
6	CASS----MYRE---------VEAFF	5
	CASSM---YREV----------EAFF	5
	CASSR---BRFY----------EQYF	8
	CASS----RERF---------YEQYF	8
7	CASSP---MRGM--------NTEAFF	78
8	CASS----IRQF---------AEQYF	180
9	CSA-----RRHI--------DNEQFF	139
10	CASSFFG-MGRM---------AEAFF	290
	CASSE---IRQM---------SPLHF	513
11	CASS----ENRI------YSNQPQHF	3
	CASSPDR-GRHM--------NTEAFF	603
	CSA-----IRDM--------QETQYF	677
12	CASSHP--RGKM--------NTEAFF	228
13	CSARD---IGRM--------GYGYTF	162
	CASS----INKI--GRLLYSNQPQHF	179
14	CASS----HREF---------GEAFF	22
	CSATRT--RRTM----------RQFF	72

FIG. 10C

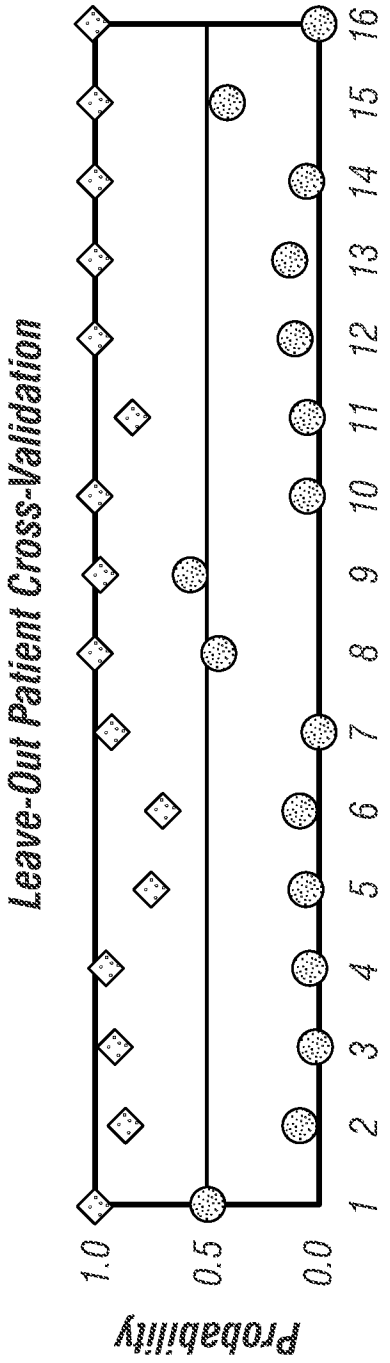


FIG. 11A

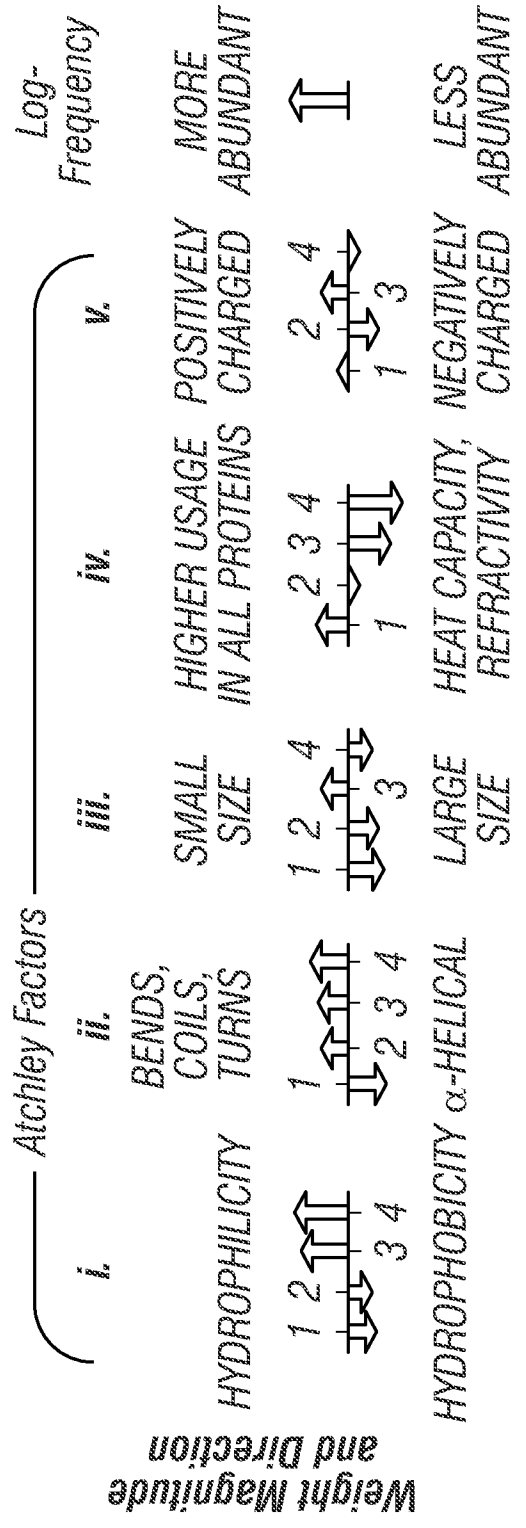


FIG. 11B

High Scoring Snippets

Patient	Snippet in CDR3	Rank
	1234	
1	CASS------LSRS----NQPQHF	2
	CASSLS-----RSNQ------PQHF	2
	CASRDS-----LSYE-------QYF	4
	CASRER-----ASYN------EQFF	14
2	CASSTGT----AGNQ------PQHF	1
3	CSVED-SGR--GSYN------EQFF	2
	CASSVTR----ASNQ------PQHF	13
4	CASSLGF----LCNN------GYTF	2
	CSVEVPTGKG-ASYE-------QYF	10
5	CASSAGIL---SSYN------EQFF	2
6	CASSP------LPRD------EQYF	6
7	CASSFDET---SSYN-----SPLHF	10
8	CASS------LDGQ---GLLGYTF	1
9	CASRRPK----PSNQ------PQHF	1
10	CAVGL------ASNE-------QFF	2
	CASSSPHRA--AYNQ------PQHF	3
	CASSSPHR---AAYN-----QPQHF	3
	CAS--------SSPHRAAYNQPQHF	3
11	CASRARRT---DSNQ------PQHF	3
12	CSVG------SSNN------EQFF	3
13	CASSQLGLAGGSSYE-------QYF	1
	CASSLEQGVG-ASNQ------PQHF	9
14	CASRQ------SSYN------EQFF	2
15	CSAGG------ASRD------IQYF	3
16	CASSQA-----SSKD------EQFF	10

FIG. 11C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/32304

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C07K 9/00, 16/00; C12Q 1/68, 1/06; G01N 33/68 (2018.01)
CPC - C07K 2317/56, 2317/565; C12Q 1/6869; G01N 33/68, 33/6857, 33/5091, 2800/00

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.
Y	ROUNDS et al. MSPrecise: A molecular diagnostic test for multiple sclerosis using next generation sequencing. Gene, 10 November 2015, Vol 572, No 2, Pages 191-197 [HHS Public Access -Author's Manuscript]. Especially abstract, pg 2 paras 2, 4, pg 3 paras 1, 3, pg 5 paras 2, 3, pg 13 fig 2b, pg 15 fig 4a,b, pg 20 table 5.	1-3, 10-15
Y	ATCHLEY et al. Solving the protein sequence metric problem. Proc Nat Acad Sci, 3 May 2005, Vol 102, No 18, Pages 6395-6400. Especially abstract, pg 6397 col 2 para 6.	1-3, 10-15
Y	WO 2016/127113 A1 (AMARANTUS BIOSCIENCE HOLDINGS) 11 August 2016 (11.08.2016). Especially claims 1-5, 34-38, sheet 4 fig 4, sheet 5 fig 5	1-3, 10-15
A	US 8,628,927 B2 (FAHAM et al.) 14 January 2014 (14.01.2014). Especially claims 1-5	1-3, 10-15
X,P	OSTMEYER et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. BMC Bioinformatics, 7 September 2017, Vol 18, page 401, (pp 1-10). Especially entire article.	1-3, 10-15

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

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
4 September 2018	26 SEP 2018

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/32304

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.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 13*ter*.1(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 13*ter*.1(a)).

      ☐ on paper or in the form of an image file (Rule 13*ter*.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:

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 18/32304

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.: Claims 4-9, 36
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III 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 1-3, 10-15

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 18/32304

Continuation of Box III: Observations where 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 1-3, 10-15, drawn to a method of identifying a disease biomarker from adaptive immune receptor sequences.

Group II+: Claims 16-25 drawn to a method of identifying a subject as having or at risk of developing the autoimmune disease, multiple sclerosis.
Group II+ will be searched upon payment of additional fee(s). The method may be searched, for example, to the extent that the B-cell heavy chain CDR3 is DFNWFD (SEQ ID NO: 1) for an additional fee and election as such. It is believed that claims 16-25 read on this exemplary invention. B-cell heavy chain CDR3s 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 B-cell heavy chain CDR3 is DFWNEV (SEQ ID NO: 5) (claims 16-25).

Group III+: Claims 26-35 drawn to a method of identifying a subject as having or at risk of developing cancer.
Group III+ will be searched upon payment of additional fee(s). The method may be searched, for example, to the extent that the cancer colorectal cancer and the T cell beta chain CDR3 is MGRM (SEQ ID NO: 53) for an additional fee and election as such. It is believed that claims 26, (28-35)(in part) read on this exemplary invention. Additional cancers and T cell beta chain CDR3s 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 cancer is breast cancer and the T cell beta chain CDR3 is SSKD (SEQ ID NO: 103) (claims 27-35)(in part)).

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 assessing 5 specific biochemical properties for each amino acid lying in the arbitrarily defined regions of the adaptive immune receptor sequences, not required by Groups II+, III+.

Croup II+ has the special technical feature of identifying one or more B-cell heavy chain CDR3 specific sequences, not required by Group I+ or III+.

Group III+ has the special technical feature of identifying one or more T-cell beta chain CDR3 specific sequences, not required by Groups I or II+.

No technical features are shared between the specific CDR3 sequences of Group II+ and/or III+ and, accordingly, these groups lack unity a priori.

Additionally, even if Groups I, II+, III+ were considered to share the technical features of:

1. obtaining the sequence of one or more heavy chain CDR3 sequences from B-cells and identifying a subject as at risk or having multiple sclerosis based on a particular CDR3 sequence.

2. obtaining the sequence of one or more T cell beta chain CDR3 sequences and identifying a subject as at risk or having colorectal cancer based on a particular CDR3 sequence.

3. obtaining the sequence of one or more T cell beta chain CDR3 sequences and identifying a subject as at risk or having breast cancer based on a particular CDR3 sequence.

These shared technical features are previously disclosed by WO 2006/116155 A2 to The Regents of the University of California (hereinafter "Univ California"), in view of the publication titled "Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue" by Sherwood et al. (hereinafter "Sherwood") [published in Cancer Immunol Immunother September 2013 Vol 62 No 9 Pages 1453-1461], in view of the publication titled "Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR" by Munson et al. (hereinafter "Munson") [published in Proc Nat Acad Sci 19 July 2016 Vol 113 No 29 Pages 8272-8277].

----continued on next sheet-----

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

As to shared technical feature #1, Univ California teaches obtaining the sequence of one or more heavy chain CDR3 sequences from B-cells and identifying a subject as at risk or having multiple sclerosis based on a particular CDR3 sequence (Pg 3 ln 15-31; "Thus, in one embodiment, this invention provides a method of diagnosing or evaluating the prognosis of multiple sclerosis (MS) or allergic encephalomyelitis (EAE) in a mammal. The method typically involves detecting the presence or quantity of an antibody in the mammal specific for a conformational epitope of myelin/oligodendrocyte glycoprotein (MOG)? the antibody specific for a conformational epitope of myelin/oligodendrocyte glycoprotein is an antibody that specifically binds to an epitope specifically bound by an antibody comprising a polypeptide sequence selected from the group consisting of SEQ ID NO:15, SEQ ID NO: 17, SEQ ID NO:19, SEQ ID NO:21, SEQ ID NO:23, SEQ ID NO:25, SEQ ID NO:27, SEQ ID NO:29, SEQ ID NO:31, SEQ ID NO:33, SEQ ID NO:35, and SEQ ID NO:37. The detecting can, optionally involve a competitive assay using a competitive binder an antibody comprising a CDR3 comprising a peptide sequence as shown in Table 2 (SEQ ID NOs: 1-12); see pg 37 Table 2 for H-CDR3 sequences").

As to share technical feature #2, Sherwood teaches obtaining the sequence of one or more T cell beta chain CDR3 sequences and identifying a subject as at risk or having colorectal cancer based on a particular CDR3 sequence (pg 9 para 2; "Given that the tumor tissue has a unique repertoire relative to the mucosal tissue, we asked if there were public, i.e. TCRB CDR3 chains shared by individuals, colorectal tumor T cell clones. Previous TCRB repertoire sequencing data found that the peripheral repertoires of unrelated individuals had an unexpectedly high number of shared TCRB CDR3 chains?.If colorectal tumor cells present similar antigens, it is possible that we could find public T cell receptors. We searched for shared TCRB CDR3 amino-acid chains between individuals and found that several individuals have a few shared TCRB chains, including a chain with the same amino-acid sequence but many different underlying DNA sequences (Table 3)"; pg 18 table 3: TCR sequences observed in multiple tumor samples [specific TCRB CDR3 sequences indicated].

As to shared technical feature #3, Munson teaches obtaining the sequence of one or more T cell beta chain CDR3 sequences and identifying a subject as at risk or having breast cancer based on a particular CDR3 sequence (pg 8275 fig 4; "Sharing of TCR pairs across breast cancer tumors reveals a shared response among HLA-A2+ patients. TCR pairs shared between seven or more patient tumors are listed (Right) with the number of patients (Left) where the specific TCR pairs were identified. A color value was assigned corresponding to the presence of the TCR in a repertoire in the tumor (blue), in the blood (red), not found (white), or not done (gray). Amino acids corresponding to the germ-line V and J sequences are underlined. Patients 1?16 are HLA-A2:01+ (HLA-A genotypes of other samples are shown in Table 1)"; specific beta chain CDR3s indicated].

As the shared technical features were known in the art at the time of the invention, they cannot be considered common special technical feature 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.

Item 4 (continued): Claims 4-9, 36 are multiple dependent claims and are not drafted according to the second and third sentences of PCT Rule 6.4(a).