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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 August 2014 (14.08.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/124451 A1

(51) International Patent Classification:
*C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/US2014/015841

(22) International Filing Date:
11 February 2014 (11.02.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/763,451 11 February 2013 (11.02.2013) US

(71) Applicant: **CB BIOTECHNOLOGIES, INC.** [US/US]; 601 Genome Way, Huntsville, AL 35806 (US).

(72) Inventors: **WANG, Chunlin**; 735 Roble Avenue #3, Menlo Park, CA 94025 (US). **HAN, Jian**; 7712 Donegal Drive, SE, Huntsville, AL 35802 (US).

(74) Agents: **CARDEN, Kellie L.** et al.; Maynard Cooper, 655 Gallatin Street, SW, Huntsville, AL 35801 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Declarations under Rule 4.17**:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

**Published**:

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

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

WO 2014/124451 A1

(54) Title: METHOD FOR EVALUATING AN IMMUNOREPERTOIRE

(57) Abstract: Disclosed is a method for amplifying RNA from T and B-cell populations and using the amplified RNA products to evaluate the possible correlation between a normal or abnormal immune response and the development of a disease such as an autoimmune disease, cancer, diabetes, or heart disease.

# METHOD FOR EVALUATING AN IMMUNOREPERTOIRE

## SEQUENCE LISTING

[0001] The instant application contains a Sequence Listing which has been submitted electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on February 5, 2014, is named 15892-0005_SL.txt and is 93,776 bytes in size.

## CROSS REFERENCE TO RELATED APPLICATION

[0002] This application is a continuation of and claims priority to U.S. Provisional Application No. 61/763,341, entitled “Method for Evaluating an Immunorepertoire” and filed on February 11, 2013, which is incorporated herein by reference.

## FIELD OF THE INVENTION

[0003] The invention relates to methods for identifying T-cell receptor antibody in a population of cells and methods for using that information to measure immune status of a patient and predict the likelihood of which disease the patient might have.

## BACKGROUND OF THE INVENTION

[0004] Scientists have known for a number of years that certain diseases are associated with particular genes or genetic mutations. Genetic causation, however, accounts for only a portion of the diseases diagnosed in humans. Many diseases appear to be linked in some way to the immune system’s response to infectious and environmental agents, but how the immune system plays a role in diseases such as cancer, Alzheimer’s, costochondritis, fibromyalgia, lupus, and other diseases is still being determined.

[0005] The human genome comprises a total number of 567-588 IG (immunoglobulin) and TR (T cell receptor) genes (339-354 IG and 228-234 TR) per

haploid genome, localized in the 7 major loci. They comprise 405-418 V, 32 D, 105-109 J and 25-29 C genes. The number of functional IG and TR genes is 321-353 per haploid genome. They comprise 187-216 V, 28 D, 86-88 J and 20-21 C genes (http://imgt.cines.fr). Through rearrangement of these genes, an estimated $2.5x10^7$ possible antibodies or T cell receptors can be generated.

[0006] A few diseases to date have been associated with the body's reaction to a common antigen (Prinz, J. et al., Eur. J. Immunol. (1999) 29(10): 3360-3368, "Selection of Conserved TCR VDJ Rearrangements in Chronic Psoriatic Plaques Indicates a Common Antigen in Psoriasis Vulgaris) and/or to specific VDJ rearrangements (Tamaru, J. et al., Blood (1994) 84(3): 708-715, "Hodgkin's Disease with a B-cell Phenotype Often Shows a VDJ Rearrangement and Somatic Mutations in the VH Genes). What is needed is a better method for evaluating changes in human immune response cells and associating those changes with specific diseases.

## SUMMARY OF THE INVENTION

[0007] The invention relates to a method for evaluating changes in immune response cell populations and associating those changes with a specific disease. In one aspect of the invention, the method comprises the steps of (a) isolating a subpopulation of white blood cells from at least one human or animal subject, (b) isolating RNA from the subpopulation of cells, (c) amplifying the RNA using RT-PCR in a first amplification reaction to produce amplicons using nested primers, at least a portion of the nested primers comprising additional nucleotides to incorporate into a resulting amplicon a binding site for a communal primer, (d) separating the amplicons from the first amplification reaction from one or more unused primers from the first amplification reaction, (e) amplifying, by the addition of communal primers in a second amplification reaction, the amplicons of the first amplification reaction having at least one binding site for a communal primer, and (f) sequencing the amplicons of the second amplification reaction to identify antibody and/or receptor rearrangements

in the subpopulation of cells. In one embodiment, the subpopulation may comprise a whole blood population or another mixed population sample.

[0008] In one embodiment, the step of isolating a subpopulation of white blood cells may be performed by flow cytometry to separate naïve B cells, mature B cells, memory B cells, naïve T cells, mature T cells, and memory T cells. In various embodiments of the method, the recombinations in the subpopulation of cells are rearrangements of B-cell immunoglobulin heavy chain (IgH), kappa and/or lambda light chains (IgK, IgL), T-cell receptor Alpha, Beta, Gamma, Delta. In an additional embodiment,

[0009] In another aspect of the invention, the method may optionally comprise an additional step comprising (g) comparing the rearrangements identified for a population of individuals to whom a vaccine has been administered with the rearrangements identified for a population of individuals to whom the vaccine was not administered to evaluate the efficacy of the vaccine in producing an immune response.

[0010] The method may also optionally comprise the additional step of (g) comparing the rearrangements identified for a population of normal individuals with the rearrangements identified for a population of individuals who have been diagnosed with a disease to determine if there is a correlation between a specific rearrangement or set of rearrangements and the disease.

[0011] In various aspects, the method can produce semi-quantitative amplification of polynucleotides comprising complementarity determining region 3 (CDR3s), which result from genetic rearrangements within T or B cells and are responsible for the affinity and specificity of antibodies and/or T cell receptors for specific antigens. Semi-quantitative amplification provides a method to not only detect the presence of specific CDR3 sequences, but also determine the relative abundance of cells which have produced the necessary recombination events to produce those CDR3 sequences.

[0012] One aspect of the invention therefore relates to a method for analyzing semi-quantitative sequence information to provide one or more immune status reports for a human or animal. The method for producing an immune status report comprising the steps of (a) identifying one or more distinct CDR3 sequences that are shared between a subject's immunoprofile and a cumulative immunoprofile from a disease library stored in a database, summing a total number of a subject's detected sequences corresponding to those shared distinct CDR3 sequences, and computing the percentage of the total number of detected sequences in the subject's immunoprofile that are representative of those distinct CDR3s shared between the subject's immunoprofile and the disease library to create one or more original sharing indices; (b) randomly selecting sequences from a public library stored in a database to form a sub-library, the sub-library comprising a number of sequences that is approximately equal to the number of distinct CDR3 sequences in the disease library, identifying one or more distinct CDR3 sequences that are shared between the subject's immunoprofile and the sub-library, summing a total number of detected sequences corresponding to those shared CDR3 sequences, and calculating a percentage of the total number of detected sequences in the subject's immunoprofile that are shared between the subject's immunoprofile and the sub-library to create a sampling sharing index; (c) repeating step (b) at least 1000 or more times; and (d) estimating the P-value as the fraction of times the sampling sharing indices are greater than or equal to the original sharing index between a patient's immunoprofile and a disease library.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The disclosure can be better understood with reference to the following drawings. The elements of the drawings are not necessarily to scale relative to each other, emphasis instead being placed upon clearly illustrating the principles of the

disclosure. Furthermore, like reference numerals designate corresponding parts throughout the several views.

[0014] Figure 1a and Figure 1b are photographs of a gel illustrating the presence of amplification products obtained by the method of the invention using primers disclosed herein.

[0015] Figure 2a and Figure 2b are cartoons representing the observed difference in diversity between an immunoprofile in an individual with a disease and an individual who is generally healthy, with each filled circle representing a distinct CDR3 sequence and the size of the circle representing the number of times that the distinct CDR3 sequence is found in the immunoprofile.

[0016] Figure 3 is a diagram illustrating the method for generating a public library.

[0017] Figure 4 is a diagram illustrating the method for generating a disease library.

[0018] Figure 5 illustrates results obtained by comparing a patient immunoprofile with a disease library, calculating a percentage for each distinct CDR3 in the patient immunoprofile that is shared between the two, and adding those percentages to produce a sum, or sharing index.

[0019] Figure 6 illustrates results obtained by comparing a patient immunoprofile with a subset of a public library, calculating a percentage for each distinct CDR3 that is shared between the two, and adding those percentages in the patient immunoprofile to produce a sum, or sharing index.

[0020] Figure 7 is a graph illustrating the method of the invention, where the area under the curve represents total sharing indices obtained for subsets of a public library (sub-libraries), a P-value is estimated, and sharing indices for comparisons of an individual's immunoprofile and one or more disease libraries are represented by vertical lines ($DL_1$, $DL_2$, etc.).

## DETAILED DESCRIPTION

[0021] The inventors have developed methods for evaluating antibody and T cell receptor rearrangements from a large number of cells, the methods being useful for comparing rearrangements identified in populations of individuals to determine whether there is a correlation between a specific rearrangement or set of rearrangements and a disease, or certain symptoms of a disease. The method is also useful for establishing a history of the immune response of an individual or individuals in response to infectious and/or environmental agents, as well as for evaluating the efficacy of vaccines.

[0022] The invention relates to a method for evaluating changes in immune response cell populations and associating those changes with a specific disease. In one aspect of the invention, the method comprises the steps of (a) isolating a subpopulation of white blood cells from at least one human or animal subject, (b) isolating RNA from the subpopulation of cells, (c) amplifying the RNA using RT-PCR in a first amplification reaction to produce amplicons using nested primers, at least a portion of the nested primers comprising additional nucleotides to incorporate into a resulting amplicon a binding site for a communal primer, (d) separating the amplicons from the first amplification reaction from one or more unused primers from the first amplification reaction, (e) amplifying, by the addition of communal primers in a second amplification reaction, the amplicons of the first amplification reaction having at least one binding site for a communal primer, and (f) sequencing the amplicons of the second amplification reaction to identify antibody and/or receptor rearrangements in the subpopulation of cells. In one embodiment, the subpopulation may comprise a whole blood population or another mixed population sample.

[0023] In one embodiment, a peripheral blood sample is taken from a patient and the step of isolating a subpopulation of white blood cells may be performed by flow cytometry to separate naïve B cells, mature B cells, memory B cells, naïve T cells, mature T cells, and memory T cells. In various embodiments of the method, the

recombinations in the subpopulation of cells are rearrangements of B-cell immunoglobulin heavy chain (IgH), kappa and/or lambda light chains (IgK, IgL), T-cell receptor Beta, Gamma, or Delta.

[0024] In a second aspect of the invention, the method may comprise an additional step (g) comparing the rearrangements identified for a population of normal individuals with the rearrangements identified for a population of individuals who have been diagnosed with a disease to determine if there is a correlation between a specific rearrangement or set of rearrangements and the disease.

[0025] In another aspect of the invention, the method may comprise an additional step comprising (g) comparing the rearrangements identified for a population of individuals to whom a vaccine has been administered with the rearrangements identified for a population of individuals to whom the vaccine was not administered to evaluate the efficacy of the vaccine in producing an immune response.

[0026] In some embodiments, the step of separating the amplicons from the first amplification reaction from one or more unused primers from the first amplification reaction may be omitted and the two amplification reactions may be performed in the same reaction tube.

[0027] The inventor previously developed a PCR method known as tem-PCR, which has been described in publication number WO2005/038039, the disclosure of which is herein incorporated by reference in its entirety. More recently, the inventor has developed a method called arm-PCR, which was described in U.S. provisional patent application number 61/042,259, the disclosure of which is herein incorporated by reference in its entirety. Also described is an apparatus for detecting target polynucleotides in a sample, the apparatus comprising a first amplification chamber for thermocycling to amplify one or more target polynucleotides to produce amplicons using nested primers, at least a portion of the nested primers comprising additional nucleotides to incorporate into a resulting amplicon a binding site for a communal primer; a means for separating the amplicons from the first amplification reaction

from one or more unused primers from the first amplification reaction; and a second amplification chamber for thermocycling to amplify one or more amplicons produced during the first amplification reaction by the addition of communal primers in a second amplification reaction, the amplicons of the first amplification reaction having at least one binding site for at least one communal primer.

[0028] Also described is a PCR chip comprising a first PCR chamber fluidly connected to both a waste reservoir and a second PCR chamber, the waste reservoir and second PCR chamber each additionally comprising at least one electrode, the electrodes comprising a means for separating amplicons produced from the first PCR chamber. The second PCR chamber is fluidly connected to a hybridization and detection chamber, the hybridization and detection chamber comprising microspheres, or beads, arranged so that the physical position of the beads is an indication of a specific target polynucleotide's presence in the sampled analyzed by means of the chip.

[0029] The tem-PCR, and especially the arm-PCR, methods provide semi-quantitative amplification of multiple polynucleotides in one reaction. Additionally, arm-PCR provides added sensitivity. Both provide the ability to amplify multiple polynucleotides in one reaction, which is beneficial in the present method because the repertoire of various T and B cells, for example, is so large. The addition of a communal primer binding site in the amplification reaction, and the subsequent amplification of target molecules using communal primers, gives a quantitative, or semi-quantitative result—making it possible to determine the relative amounts of the cells comprising various rearrangements within a patient blood sample. Clonal expansion due to recognition of antigen results in a larger population of cells which recognize that antigen, and evaluating cells by their relative numbers provides a method for determining whether an antigen exposure has influenced expansion of antibody-producing B cells or receptor-bearing T cells. This is helpful for evaluating whether there may be a particular population of cells that is prevalent in individuals

who have been diagnosed with a particular disease, for example, and may be especially helpful in evaluating whether or not a vaccine has achieved the desired immune response in individuals to whom the vaccine has been given.

[0030] There are several commercially available high throughput sequencing technologies, such as Roche Life Sciences's 454 sequencing. In the 454 sequencing method, 454A and 454B primers are linked onto PCR products either during PCR or ligated on after the PCR reaction. When done in conjunction with tem-PCR or arm-PCR, 454A and 454B primers may be used as communal primers in the amplification reactions. PCR products, usually a mixture of different sequences, are diluted to about 200 copies per µl. In an "emulsion PCR" reaction, (a semisolid gel like environment) the diluted PCR products are amplified by primers (454A or 454B) on the surface of the microbeads. Because the PCR templates are so dilute, usually only one bead is adjacent to one template, and confined in the semisolid environment, amplification only occurs on and around the beads. The beads are then eluted and put onto a plate with specially designed wells. Each well can only hold one bead. Reagents are then added into the wells to carry out pyrosequencing. A fiber-optic detector may be used to read the sequencing reaction from each well and the data is collected in parallel by a computer. One such high throughput reaction could generate up to 60 million reads (60 million beads) and each read can generate about 300bp sequences.

[0031] One aspect of the invention involves the development of a database of "personal immunorepertoires," or immunoprofiles, so that each individual may establish a baseline and follow the development of immune responses to antigens, both known and unknown, over a period of years. This information may, if information is gathered from a large number of individuals, provide an epidemiological database that will produce valuable information, particularly in regard to the development of those diseases, such as cancer and heart disease, which are thought to often arise from exposure to viral or other infectious agents or transformed

cells, many of which have as yet been unidentified. One particularly important use for the method of the invention involves the evaluation of children to determine whether infectious disease, environmental agents, or vaccines may be the cause of autism. For example, many have postulated that vaccine administration may trigger the development of autism. However, many also attribute that potential correlation to the use of agents such as thimerosol in the vaccine, and studies have demonstrated that thimerosol does not appear to be a causative agent of the disease. There is still speculation that the development of cocktail vaccines has correlated with the rise in the number of cases of autism, however, gathering data to evaluate a potential causal connection for multiple antigens is extremely difficult. The method of the present invention simplifies that process and may provide key information for a better understanding of autism and other diseases in which the immune response of different individuals may provide an explanation for the differential development of disease in some individuals exposed to an agent or a group of agents, while others similarly exposed do not develop the disease.

[0032] Imbalances of the immunoprofile, triggered by infection, may lead to many diseases, including cancers, leukemia, neuronal diseases (Alzheimer's, Multiple Sclerosis, Parkinson's, autism etc.), autoimmune diseases, and metabolic diseases. These diseases may be called immunoprofile diseases. There may be two immunoprofile disease forms. (1) a "loss of function" form, and (2) a "gain of function" form. In the "loss of function" form, a person is susceptible to a disease because his/her restricted and/or limited immunoprofile lacks the cells that produce the most efficient and necessary IGs and TRs. In the "gain of function" form, a person is susceptible to a disease because his/her immunoprofile gained cells that produce IGs and TRs that normally should not be there. In the "loss of a function" (LOF) immunoprofile diseases, an individual does not have the appropriate functional B or T cells to fight a disease. His/her HLA typing has determined that those cells are

eliminated during the early stages of the immune cell maturation process, the cells generally being eliminated because they react to strongly to his/her own proteins.

[0033] One aspect of the invention also provides a method comprising (a) amplifying and sequencing one or more RNAs from the T cells and/or B cells from one or more individuals, (b) inputting the sequences into a database to provide data which may be stored on a computer, server, or other electronic storage device, (c) inputting identifying information and characteristics for an individual corresponding to the sequences of the one or more RNAs as data which may also be stored on a computer, server, or other electronic storage device, and (d) evaluating the data of step (b) and step (e) for one or more individuals to determine whether a correlation exists between the one or more RNA sequences and one or more characteristics of the individual corresponding to the sequence(s). Identifying information may include, for example, a patient identification number, a code comprising the patient's HLA type, a disease code comprising one or more clinical diagnoses that may have been made, a "staging code" comprising the date of the sample, a cell type code comprising the type of cell subpopulation from which the RNA was amplified and sequenced, and one or more sequence codes comprising the sequences identified for the sample.

[0034] The described method includes a novel primer design that not only allows amplification of the entire immunorepertoire, but also allows amplification in a highly multiplex fashion and semiquantitatively. Multiplex amplification requires that only a few PCR or RT-PCR reactions will be needed. For example, all IGs may be amplified in one reaction, or it could be divided into two or three reactions for IgH, IgL or IgK. Similarly, the T-cell receptors (TRs) may be amplified in just one reaction, or may be amplified in a few reactions including TRA, TRB, TRD, and TRG. Semi-quantitative amplification means that all the targets in the multiplex reaction will be amplified independently, so that the end point analysis of the amplified products will reflect the original internal ratio among the targets.

[0035] In various aspects, the method can produce semi-quantitative amplification of polynucleotides comprising complementarity determining regions (CDRs), which result from genetic rearrangements within T or B cells and are responsible for the affinity and specificity of antibodies and/or T cell receptors for specific antigens. Semi-quantitative amplification provides a method to not only detect the presence of specific CDR3 sequences, but also determine the relative numbers of cells which have produced the necessary recombination events to produce those CDR3 sequences.

[0036] One aspect of the invention therefore relates to a method for analyzing semi-quantitative sequence information to provide one or more immune status reports for a human or animal. The method for producing an immune status report comprising the steps of (a) identifying one or more distinct CDR3 sequences that are shared between a subject's immunoprofile and a disease library stored in a database, summing the total of those shared CDR3 sequences and computing the percentage of the total number of sequences in the subject's immunoprofile that are shared between the subject's immunoprofile and the disease library to create one or more original sharing indices; (b) randomly selecting sequences from a public library stored in a database to form a sub-library, the sub-library comprising a number of sequences that is approximately equal to the number of distinct sequences in the disease library, identifying one or more distinct CDR3 sequences that are shared between the subject's immunoprofile and the sub-library, summing the total of those shared CDR3 sequences and calculating the percentage of the total number of sequences in the subject's immunoprofile that are shared between the subject's immunoprofile and the sub-library to create a sampling sharing index; (c) repeating step (b) at least 1000 or more times; and (d) estimating the P-value as the fraction of times the sampling sharing indices are greater than or equal to the original sharing index between a patient's immunoprofile and a disease library.

[0037] The inventors have discovered that the immunoprofile of individuals who have certain diseases, such as, for example, cancer, autoimmune disease, etc., may be characterized by a lack of diversity in one or more immune cell population(s). Figure 1 is a cartoon illustrating the difference that may be observed between, for example, the distinct type and number of T-cells present in a blood sample from a cancer patient (Fig. 1a) and a healthy patient (Fig. 1b), where each circle represents a distinct type of T-cell, as represented by an amplified and sequenced recombined cDNA of the complementarity determining region of the T-cell receptor (e.g., CDR3), and the relative number of cells which are determined, by PCR amplification and sequencing, to share the same CDR3 sequence. As Fig. 1a indicates, there may be fewer distinct cells of different specificities, but larger numbers of cells of certain specificities, as represented by the CDR3 sequences. Fig. 1b illustrates a normal profile of more different cells, but fewer numbers of each type of cell sharing the same CDR3 sequence.

[0038] The list of each distinct CDR3-expressing cell, and the numbers of such cells represented within a blood or tissue sample from a human or animal, can constitute an immunoprofile for that human or animal. Compiling the immunoprofiles from a group of humans, for example, the group comprising both healthy individuals and individuals with various different diseases may provide a “public library” that is representative of the type of diversity found in a normal population (Fig. 2). Similarly, compiling the immunoprofiles of a group of individuals who have been clinically diagnosed with a particular disease may provide a “disease library” that is representative of the lack of diversity, the specific CDR3s of the expanded populations of cells, etc. (Fig. 3). These immunoprofiles may be stored in a database, accessible via computer access to the internet, for example, so that the information may be used in the method of the invention to analyze the immune status of a patient.

[0039] An immunoprofile, comprising a listing of distinct CDR3-expressing cells (“distinct CDR3s”, those cells sharing a unique CDR3 sequence) and the numbers of each distinct CDR3 present in a blood or tissue sample from an individual may be

produced for an individual patient. The patient's immunoprofile is compared to the combined immunoprofiles of a group of patients who have been diagnosed with a particular disease (a disease library, stored in a database). This can be done for a series of disease libraries, and shown in Fig. 4.

[0040] Millions of possible combinations are possible for the public library, the immune systems of most of those individuals generally exhibiting increased diversity over that of a group of individuals who have been diagnosed with a specific disease. Therefore, the inventors determined that an accurate assessment and comparison for the method of the invention would be facilitated by the step of preparing sub-libraries by randomly sampling/selecting from the lists of distinct CDR3s and their numbers in the public library. The number of distinct CDR3s, represented by unique peptide sequence of CDR3 fragments, should be approximately equal to the number of distinct CDR3s identified in the disease library, or an average calculated from more than one disease library. Producing a significant number of sub-libraries, such as, for example, 1000 or more sub-libraries, produced by randomly sampling from the public library, increases the presence of a variety of distinct CDR3s and produces a result that is statistically significant effective for identifying and characterizing an individual patient's immunoprofile as normal ("healthy") or characterized by the presence of a type and number of cells that have been associated with a particular disease.

[0041] In the method of the invention, a patient supplies a clinical sample comprising, for example, blood or tissue, from which distinct CDR3s are semi-quantitatively amplified and sequenced. This provides the identity and the relative abundance of each CDR3 for all distinct CDR3s. This information may be entered into a program which accesses a database containing at least one public library and one or more disease libraries. Software used for data entry and/or analysis may be accessed via internet access to the database, or may be located on an individual personal computer, with internet access to the sequence information in the database. Comparisons are obtained between the individual immunoprofile and the various libraries and sub-

libraries, and results are generated as generally illustrated in Fig. 4 and Fig. 5, where specific CDR3 sequences are detected, the numbers of those distinct CDR3 sequences detected are counted, and a determination is made as to whether or not that specific distinct CDR3 is present in both the individual's immunoprofile and a specific library (i.e., that specific distinct CDR3 is "shared" between the individual and the library). The percentages representing numbers of those CDR3s that are determined to be shared are added together to produce a sum comprising the fraction of the total that comprises CDR3s in the individual's immunoprofile shared between the individual's immunoprofile and the specific library (i.e., a "sharing index"). From the results obtained for the sub-libraries, a P-value is calculated as the probability that a random percentage would be greater than or equal to the percentage noted for a particular disease library, and a significant result is noted when the fraction of times the sampling sharing indices exceeds the original sharing index for a particular library is less than 0.01, for instance. If that sharing index represents the relationship between the individual's immunoprofile and a disease library, the individual may then be informed of the likelihood that the individual/patient has the disease represented by the specific disease library. If P-values computed against all disease libraries is greater than 0.01, the individual's report may indicate that the immune profile looks normal and the disease state has not been detected.

[0042] As sequence data is compiled and stored in one or more databases for multiple populations of individuals, it may additionally be possible to associate certain sharing indexes with libraries representing populations with pre-conditions or predispositions to certain diseases. The immune system is both proactive and reactive, and changes in the immune system, reflected in the immunoprofile, may provide the first—and sometimes the only—signal that a predisposition, a precondition, or even an established disease is present. The inventors have utilized the method to demonstrate that certain types of cancers, inflammatory bowel disease, and certain viral infections may be detected by determining the sharing index between a patient and an established

disease library, obtained by sequencing CDR3s using the ARM-PCR method to produce a subset of the immunorepertoire representing the CDR3s present.

[0043] The results are even more reliable when a filter is applied to the sequence data. For example, the inventors have developed a "SMART" filter for the sequence data that aids in the generation of significantly more reliable results. This is described further in the Examples.

[0044] By way of further explanation, the following example may be illustrative of the methods of the invention. Blood samples may be taken from children prior to administration of any vaccines, those blood samples for each child establishing a "baseline" from which future samples may be evaluated. For each child, the future samples may be utilized to determine whether there has been an exposure to an agent which has expanded a population of cells known to be correlated with a disease, and this may serve as a "marker" for the risk of development of the disease in the future. Individuals so identified may then be more closely monitored so that early detection is possible, and any available treatment options may be provided at an earlier stage in the disease process.

[0045] By means of providing another example, blood samples may be taken from children prior to administration of any vaccines, those blood samples from each child establishing a "baseline" from which future samples may be evaluated. For each child and for the entire population of children in the study, those baselines may be compared to the results of RNA sequencing of T and B cells using target-specific primers to amplify antibody and T-cell receptor, after vaccine administration. The comparison may further involve the evaluation of data regarding symptoms, diagnosed diseases, and other information associated for each individual with the corresponding antibody, and T-cell receptor sequences. If a relationship exists between the administration of a vaccine and the development of a particular disease, individuals who exhibit symptoms of that disease may also share a corresponding antibody or T-cell receptor, for example, or a set of corresponding antibodies or T-cell receptors.

[0046] The method of the invention may be especially useful for identifying commonalities between individuals with autoimmune diseases, for example, and may provide epidemiological data that will better describe the correlation between infectious and environmental factors and diseases such as heart disease, atherosclerosis, diabetes, and cancer—providing "biomarkers" that signal either the presence of a disease, or the tendency to develop disease.

[0047] The method may also be useful for development passive immunity therapies. For example, following exposure to an infectious agent, certain antibody-producing B cells and/or T cells are expanded. The method of the invention enables the identification of protective antibodies, for example, and those antibodies may be utilized to provide passive immunity therapies in situations where such therapy is needed.

[0048] The method of the invention may also provide the ability to accomplish targeted removal of cells with undesirable rearrangements, the method providing a means by which such cells rearrangements may be identified.

[0049] The inventor has identified and developed target-specific primers for use in the method of the invention. T-cell-specific primers are shown in Table 1, and antibody-specific primers are shown in Table 2. An additional embodiment of the invention is a method of using any one or a combination of primers of Table 1 or Table 2, to amplify RNA from a blood sample, and more particularly to identify antibodies, T-cell receptors, and HLA molecules within a population of cells.

[0050] Arm-PCR or tem-PCR may be used to amplify genes coding for the immunoglobulin superfamily molecules in am amplification method described previously by the inventor (Han et al., 2006. Simultaneous Amplification and Identification of 25 Human Papillomavirus Types with Templex Technology. J. Clin. Micro. 44(11). 4157-4162). In a tem-PCR reaction, nested gene-specific primers are designed to enrich the targets during initial PCR cycling. Later, universal "Super" primers are used to amplify all targets. Primers are designated as $F_o$ (forward out), $F_i$ (forward in), $R_i$ (reverse in), $R_o$ (reverse out), FS (forward super primer) and RS,(reverse super primer), with super

primers being common to a variety of the molecules due to the addition of a binding site for those primers at the end of a target-specific primer. The gene-specific primers ($F_o$, $F_i$, $R_i$, and $R_o$) are used at extremely low concentrations. Different primers are involved in the tem-PCR process at each of the three major stages. First, at the "enrichment" stage, low-concentration gene-specific primers are given enough time to find the templates. For each intended target, depending on which primers are used, four possible products may be generated: $F_o/R_o$, $F_i/R_o$, $F_i/R_i$, and $F_o/R_i$. The enrichment stage is typically carried out for 10 cycles. In the second, or "tagging" stage, the annealing temperature is raised to 72°C, and only the long 40-nucleotide inside primers ($F_i$ and $R_i$) will work. After 10 cycles of this tagging stage, all PCR products are "tagged" with the universal super primer sequences. Then, at the third "amplification" stage, high-concentration super primers work efficiently to amplify all targets and label the PCR products with biotin during the process. Specific probes may be covalently linked with Luminex color-coated beads.

[0051] To amplify the genes coding for immunoglobulin superfamily molecules, the inventor designed nested primers based on sequence information in the public domain. For studying B and T cell VDJ rearrangement, the inventor designed primers to amplify rearranged and expressed RNAs. Generally, a pair of nested forward primers is designed from the V genes and a set of reverse nested primers are designed from the J or C genes. The average amplicon size is 250-350bp. For the IgHV genes, for example, there are 123 genes that can be classified into 7 different families, and the present primers are designed to be family specific. However, if sequencing the amplified cDNA sequences, there are enough sequence diversities to allow further differentiation among the gene within the same family. For the MHC gene locus, the intent is to amplify genomic DNA.

## EXAMPLES

*Calculation of Sharing Index*

[0052] Assuming that **S** is a subject's immunoprofile (IP), which is represented by **N** unique CDR3 sequences $CDR3_1$, $CDR3_2$, … $CDR3_n$, each CDR3 has its own frequency $s_1$, $s_2$, … $s_n$.

[0053] **D** is a disease library, which is the sum of a certain number of patients' immunoprofile with **M** unique CDR3s. All patients in the disease library were diagnosed to have the same disease.

[0054] **P** is a public library, which is the sum of a large number of control's immunoprofile.

[0055] The **S**haring **I**ndex is defined as the sum of $s_x$, $s_y$, … $s_z$, where $CDR3_x$, $CDR3_y$, … $CDR3_z$ are shared in the subject's immunoprofile and a library. Note that $s_x$, $s_y$, … $s_z$ is the frequency of CDR3s in the subject's immunoprofile, not in the library.

[0056] Assuming that there are always more unique CDR3s in a public library (P) than in a disease library (D), **M** unique CDR3s in the public library are randomly selected and used to create a sub-library P1 and the sharing index ($SI_{p1}$) between the subject and the sub-library computed according to above formula. The sampling procedure is repeated 1000 or more times and 1000 or more $SI_{px}$ are computed.

[0057] The sharing index $SI_d$ between the subject and the disease library are computed in the same manner. The **P-value** is defined as the fraction of all **SI**s ($SI_{p1}$, $SI_{p2}$, … $SI_{px}$, $SI_d$, (Note that $SI_d$ is included), which is equal to or greater than $SI_d$. Note that, when sampling CDR3s in the public library, CDR3s found in **x** control's immunoprofiles are given **x** times of chances to be sampled.

*Amplification of T or B Cell Rearrangement Sites*

[0058] All oligos were resuspended using 1x TE. All oligos except 454A and 454B were resuspended to a concentration of 100pmol/µL. 454A and 454B were resuspended to a concentration of 1000pmol/µL 454A and 454B are functionally the same as the communal primers described previously, the different sequences were used for follow up high throughput sequencing procedures.

[0059] Three different primer mixes were made. An Alpha Delta primer mix included 82 primers (all of TRAV-C + TRDV-C), a Beta Gamma primer mix included 79 primers (all of TRBVC and TRGV-C) and a B cell primer mix that included a total of 70 primers. $F_o$, $F_i$, and $R_i$ primers were at a concentration of 1pmol/µL. $R_o$ primers were at a concentration of 5 pmol/µL. 454A and 454B were at a concentration of 30 pmol/µL.

[0060] Three different RNA samples were ordered from ALLCELLS (www.allcells.com). All samples were diluted down to a final concentration of 4 ng/uL. The samples ordered were:

| **Cell type:** | **Source:** |
|---|---|
| ALL-PB-MNC | A patient with acute lymphoblastic leukemia |
| NPB-Pan T Cells | Normal T cells |
| NPB-B Cells | Normal B cells |

[0061] RT-PCR was performed using a Qiagen One-Step RT-PCR kit. Each sample contained the following:

10 µL of Qiagen Buffer
2 µL of DNTP's
2 µl of Enzyme
23.5 µL of $dH_2O$
10 µL of the appropriate primer mix
2.5 µL of the appropriate template (10ng of RNA total)

The samples were run using the following cycling conditions:

50°C for 30 minutes

95°C for 15 minutes
94°C for 30 seconds
15 cycles of
55°C for 1 minute
72°C for 1 minute
94°C for 15 seconds
6 cycles of
70°C for 1 minute 30 seconds
94°C for 15 seconds
30 cycles of
55°C for 15 seconds
72°C for 15 seconds
72°C for 3 minutes
4°C Hold

[0062] The order of samples placed in the gel shown in Fig. 1a was: (1) Ladder (500bp being the largest working down in steps of 20bp, the middle bright band in Fig. 1a is 200bp); (2) α + δ primer mix with 10ng Pan T Cells Template; (3) β + γ primer mix with 10ng Pan T Cells Template; (4) B Cell primer mix with 10ng B Cells Template; (5) B Cell primer mix with 10ng ALL Cells Template; (6) α + δ primer mix with 10ng ALL Cells Template; (7) β + γ primer mix with 10ng ALL Cells Template; 8. α + δ primer mix blank; (9) β + γ primer mix blank; (10) B Cell primer mix blank; (11)Running buffer blank. These samples were run on a pre-cast ClearPAGE® SDS 10% gel using 1X ClearPAGE® DNA native running buffer.

[0063] The initial experiment showed that a smear is generated from PCR reactions where templates were included. The smears indicate different sizes of PCR products were generated that represented a mixture of different VDJ rearrangements. There is some background amplification from the B cell reaction. Further improvement on that primer mix was required to clean up the reaction.

[0064] To determine whether the PCR products indeed include different VDJ rearrangements, it was necessary to isolate and sequence the single clones. Instead of using the routine cloning procedures, the inventor used a different strategy. PCR products generated from the Alpha Delta mix and the Beta Gamma mix (lanes 2 and 3 in Fig. 1a) were diluted 1:1000 and a 2 µl aliquot used as PCR template in the following reaction. Then, instead of using a mixture of primers that targeting the entire repertoire,

one pair of specific Fi and Ri primers were used (5 pmol each) to amplify only one specific PCR product. The following cycling conditions were used to amplify the samples:

95°C for 5 minutes
30 cycles of
94°C for 30 seconds
72°C for 1 minute
72°C for 3 minutes
4°C hold

[0065] A Qiagen PCR kit was used to amplify the products. The Master Mix used for the PCR contained the following:

| | Per Reaction | Master Mix x 12 |
|---|---|---|
| 10x PCR Buffer | 5µL | 60µL |
| dNTP | 1µL | 12µL |
| HotStartTaq Plus | 0.25µL | 3µL |
| $H_2O$ | 39.75 µL | 477 µL |

[0066] The photograph of the gel in Fig. 1b shows the PCR products of the following reactions: (1) Ladder; (2) TRAV1Fi+TRACRi with alpha delta Pan T PCR product; (3) TRAV2Fi+TRACRi with alpha delta Pan T PCR product; (4) $TRAV3F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (5) $TRAV4F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (6) $TRAV5F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (7) $TRAV1F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (8) $TRAV2F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (9) $TRAV3F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (10) $TRAV4F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (11) $TRAV5F_i$+$TRACR_i$ with alpha delta Pan T PCR product; (12) PCR Blank. Primers listed as $F_1$ are "forward inner" primers and primers listed as $F_o$ are "forward outer" primers, with $R_i$ and $R_o$ indicating "reverse inner" and "reverse outer" primers, respectively.

[0067] As illustrated by Fig. 1b, a single PCR product was generated from each reaction. Different size bands were generated from different reactions. This PCR cloning approach is successful for two major reasons--(1) The PCR templates used in this reaction were diluted PCR products (1:1000) of previous reactions that used primer mixes to amplify all possible VDJ rearrangements (for example, a primer mix was used that included total of 82 primers to amplify T cell receptor Alpha and Delta genes) and (2) Only one pair of PCR primers, targeting a specific V gene, are used in each reaction during this "cloning" experiment. Some of these products were gel purified and sequenced. The following are example sequences obtained from the protocol described above. In every case, a single clone was obtained, and a specific T cell receptor V gene that matched the Fi primer was identified.

TRAV1 template + 454A as sequencing primer:
NNNNNNNNNNCNTANTCGGTCTAAGGGTACNGNTACCTCCTTTTGAAGGAGCT
CCAGATGAAAGACTCTGCCTCTTACCTCTGTGCTGTGAGAGATANCAACNATCA
CTTAATCTTGGGCGCTGGGAGCAGACTAATTATAATGCCAGATATCCACAACCC
TGACCCTGCCGCGTACCAGCTGAAAGACTATGAACAGGATGGGGAGGCAGNAG
NAGNAG (SEQ ID NO. 1)

TRAV1 template + 454A as sequencing primer:
NNNNNNNNNNGNANGNNCAGGGTTCTGGATATTTGGTTTNACAATTAGCTTGGT
CCCTGCTCCAAAGATTAATTTGTAGTTGCTATCCCTCACAGCACAGAGGTAAGA
GGAAGAGTATTTCTTCTGGAGCTCCTTCAACAGGAGGAAACTGTACCCTTTATA
CCTACTAAGGAATGAAGA (SEQ ID NO. 2)

TRAV2 template + 454A as sequencing primer:
NNNNNNNNNNNNTNNCGGTTCTCTTNNTCGCTGCTCATCCTCCAGGTGCGGGA
GGCAGATGCTGCTGTTTACTACTGTGCTGTGNANNANGGCANNGACAACAACCT
CNTCTTTGGTGGAGGNACCCTACTNNTGGTTATNCCNAATANCCANAACCCTGA
CCCTGCCGAGNAGCAGCANAAAAACTNNNAGGGGGGTGGAGAAGNANNNNN
(SEQ ID NO. 3)

TRAV3 template + 454A as sequencing primer:
NNNNNNNNNNNNNNGGNNNGGNAGCTATGGCTTTGAAGCTGAATTTAACAAGA
GCCAAACCTCCTTCCACCTGAAGAAACCATCTGCCCTTGTGAGCGACTCCGCTT
TGTACTTCTGTGCTGTGAGAGACATCAACGCTGCCGGCAACAACCTAACTTTTG
GAGGAAGAACCATGGTGCTAGTTAAACCAAATATCCATAACCCTGACGCTGCCG
TGTACCAGCTGAAAGACTCTGAGGGGGCTGGAGAGGNAGGNG (SEQ ID NO. 4)

TRAV4 template + 454A as sequencing primer:
NNNNNANNGGNNNNNGTTTATCCCTGCCGACAGAAAGTCCAGCACTCTGAGCC
TGCCCCGGGTTTCCCTGAGCGACACTGCTGTGTACTACTGCCTCGTGGGTGAC
CGGTCTGGAAACAGCGATGAAATTTTCATCTTAGGAAGAAGAACGCTTCTAGTC
ATCCANCCCAACATCCACAACCCTGCCGCGGAGNAGCACCAGAAAAAAGATGA
TGAGGGGGANGNAGNAGNANNNN (SEQ ID NO. 5)

TRAV5 template + 454A as sequencing primer:
NNNNNNNNNNNNNNNNTCNCTGNTCTATTGAATAAAAAGGATAAACATCTGTCT
CTGCGCATTGCAGACACCCAGACTGGGGACTCAGCTATCTACTTCTGTGCAGA
GAGCCCCGGTGGCGGCAGCAACTTCTTCTTTGGTGGAGGAGCANTACTACTAG
TCGTTCTACATANCCACAACCATGATNCCGCCGAGTACNTGCTGAAAAAATATG
ATGAGGATGGAGAAGAAGNAGCATNAN (SEQ ID NO. 6)

TRBV19Fi template + 454A as sequencing primer:
NNNNNNNNCTGAGGGTANNCGTCTCTCGGGAGAAGAAGGAATCCTTTCCTCTC
ACTGTGACATCGGCCCAAAAGAACCCGACAGCTTTCTATCTCTGTGCCAGTAGT
ATGGGGGGGGGGGCCTACAATGAGNACGGCGGCGGGGGAGGGACNNTGCTC
GTCGTGGAGGAGGACATGAAGGTCTTGCCCGCNNCNGAGGAAGNTGNANANG
AACCATAAAAATGCGCTGGCTGAANNN (SEQ ID NO. 7)

TRBV20Fi template + 454A as sequencing primer:
NNNNNNNNNNNGCTCNNNNNNCNCATACGAGCAAGGCGTCGAGAAGGACAAG
TTTCTCACAACCATGCAAGCCTGACCTTGTCCACTCTGACAGTGACCAGTGCCC
ATCCTGAAGACAGCAGCTTCTACATCTGCAGTGCTAGAGGGGGGGGGGGGGA
CGACTACTACTACTTCGGCGGGGGGGGCATGCTGATCGTGGAGGAGGAGGAC
ATGNAGCTCCTCCCCGCCGCCGAGGTTGTTGTGTNTNNANCATCATACTGNTG
GTGGAGNAGNAGNAGCN (SEQ ID NO. 8)

TRBV21Fi template + 454A as sequencing primer:
NNNNNNNNNNNNNNNGNNNNNNNNNNNTACTTTCNGAATGAAGAACTTATTCA
GAAAGCAGAAATAATCAATGAGCGATTTTTAGCCCAATGCTCCAAAAACTCATCC
TGTACCTTGGAGTTCCAGTCCACGGAGTCAGGGGACACAGCACTGTATTTCTGT
GCCAGCAGCA (SEQ ID NO. 9)

TRBV23Fi template + 454A as sequencing primer:
NNNGNNNNNNNANNGGANANGCACAAGAAGCGATTCTCATCTCAATGCCCCAA
GAACGCACCCTGCAGCCTGGCAATCCTGTCCTCAGAACCGGGAGACACGGCAC
TGTATCTCTGCGCCAGCAGTCAATCGGGGGGGGGGGGGAGGGCCGTCCGCAG
CGGGGGGGGGGGGGGCCGGGGGACGGTCCCAAAGAGAAAGAAAACCTGCCC
CCCGCGCTCGGGCGGTGTGATTGAGCGAAACAGACAGGAAGGNAAGNAAAAAA
NNNNANCNNCNCTCNN (SEQ ID NO. 10)

TRBV24Fi template + 454A as sequencing primer:
NNNNNNNNGNNANNNTCTGATGGANACAGTGTCTCTCGACAGGCACAGGCTAA
ATTCTCCCTGTCCCTAGAGTCTGCCATCCCCAACCAGACAGCTCTTTACTTCTGT
GCCACCAGTGANGCGGGGGGCGGGGACCACTACTTCGGGGGGGGGAGGCGG
ACCAGGGTGCTGGTCGACGAGAAAAAGGAGCTCCCCCCCGCCGCCGCTGTGG
TTGTTGCTTCATAATAATCAGGNNGGNGAGGNAGNAGNAANN (SEQ ID NO. 11)

[0068] To investigate the impact of artifacts on the overall repertoire analysis of the TCRβ transcriptome, the inventors conducted control experiments using chemically synthesized TCRβ CDR3 templates. For this, the inventors chemically synthesized four distinct clones, clonally purified each clone, and prepared different mixes of the four constructs as templates for amplicon rescue multiplex (ARM)-PCR. Two different

reaction mixtures were subjected to two independent ARM-PCR reactions, and the pooled PCR products were sequenced at a length of 100bp from both ends using the Illimuna HiSeq2000®. The inventors first joined together paired-end reads through overlapping alignment with a modified Needleman-Wunsch algorithm, and then mapped the merged sequences to germline V, D and J reference sequences.

[0069] Without cleaning, the inventors obtained a total of 5,729,613 sequences from template mix I that could be mapped to TCRβ V, D and J segments. Surprisingly, the sequence reads purportedly represented a total of 36,439 unique CDR3 variants. Therefore, given that only four distinct CDR3 variants were present in the template mixtures, virtually all of the identified CDR3 variants must be non-authentic. Similar results were obtained for the second template mix, in which a total of 9,131,681 VDJ-mapped sequences were identified that mimicked the existence of 50,354 unique TCRβ CDR3 variants. The inventors' independent sequencing experiments show that only a few distinct CDR3 template variants can create artifactual repertoire diversities that far outweigh the real template diversity, and thus the inventors set out to eliminate these artifacts.

[0070] The quality of 3' end Illumina sequencing reads is generally considered to be low. In the context of repertoire sequencing, this is troublesome because PCR primers need to be positioned distal enough from the hypervariable V(D)J junctions to avoid negative effects due to primer-template mismatching. As a consequence, the CDR3 segments of interest are generally "shifted" closer to the 3' end of the sequencing reads, the region with increased sequencing error rates. Another technical issue that deserves attention is the observation that sequencing errors are context-specific and consequently strand-specific. Therefore, it is realistic to assume that the probability that a sequencing error in a forward read coincides with that in the corresponding reverse read is rare.

[0071] Considering this, the inventors devised a paired-end strategy that affords double-strand sequencing of complete TCR CDR3 segments on the basis of the

Illumina® technology. In this approach, forward and reverse sequencing primers are positioned at the framework region 3 and at the TCR J region or the 5' end of the C region, respectively. Taking into account the average length of Illumina sequence reads (currently 100-150 bp) this design enables the complete sequencing of both strands that define a CDR3 segment. In a second step, the forward and reverse reads are then analyzed for sequence mismatches and CDR3 sequences that exhibit non-identity of both strands are eliminated using a newly developed paired-end filtering algorithm.

[0072] Applying this sequencing error filter to the 5,729,613 CDR3 sequences obtained for template mix I, the inventors identified a total of 2,751,131 (48%) CDR3 sequences that contained conflicting sequence information on their opposite strands. Discarding of these sequences resulted in the elimination of 35,455 (97.2%) distinct artifactual CDR3 variants. Consistent with this, the paired-end filter removed 4,308,020 (47%) CDR3 sequences from template mix II, leading to the elimination of 49,063 (97.4%) artifactual CDR3 variants. A total of 973 and 1271 unique CDR3 variants, respectively, passed through the filter. These results indicate that paired-end sequencing and filtering reduces the total number of non-authentic unique CDR3 sequences by almost two orders of magnitude.

[0073] Detailed analysis of the frequency distribution of the non-authentic CDR3 variants after the sequencing error filter revealed that in both mixtures approximately 50% of all artifacts were single-copy sequences. About 10% of these artifactual CDR3s displayed >100 copy numbers and accounted for > 80% of all artifactual CDR3 variants. Given that variable TCR genes do not undergo somatic hypermutation, the inventors developed a reference algorithm that identifies and removes CDR3 sequence reads that display nucleotide mismatches relative to the mapped germline V, D and J reference sequences, as these must be artifacts generated at the level of PCR amplification or sequencing.

[0074] Applying this filtering algorithm to the "paired-end filtered" sequences of template mix I, a total of 29,804 sequences, which corresponded to 609 unique CDR3

variants, were removed. For template mix II, 54,516 artifactual sequences (831 unique CDR3 variants) were identified. Thus, the use of the reference sequence filter leads to a 60% reduction of non-authentic distinct CDR3 sequences. The reference filter is ineffective at the V-J and D-J junctions because the randomly added nucleotides in these regions during somatic recombination cannot be mapped. Therefore, the inventors implemented a PCR filter after computational simulation experiments to better understand four variables: the impact of the initial template number, the replication efficiency of each cycle, the cycle number (n), and the DNA polymerase error rate (μ) on the total end-point error rate. In contrast, the inventors noted that the PCR polymerase error rate has a pronounced effect on the number of accumulated errors

[0075] In the inventors' control sequencing experiments, PCR amplification was performed with 15 cycles and 45 cycles in the first and second reaction, using Taq polymerase. To simulate error accumulation during the ARM-PCR reactions more realistically, the PCR efficiency was set to decreased 5% per cycle for the first 25 cycles and 10% per cycle for the remaining cycles. The PCR efficiency was reset to 1.0 for each fresh PCR reaction. Furthermore, the inventors allowed mutation at the second position. Published substitution error rates for Taq enzyme, expressed as errors per bp per cycle, range from $0.023 \times 10^{-4}$ to $2.1 \times 10^{-4}$. In the simulation experiments, the substitution error rate was set at $2.7 \times 10^{-5}$, and the insertion-deletion (indel) error rate was set as $1.0 \times 10^{-6}$. Taq polymerase is known to have a much higher insertion-and-deletion (indel) mutation rate in homopolymeric region of templates. For a homopolymeric region, indel mutation in any position of this region generates identical pattern. Therefore, the indel error rate in a homopolymeric region was set as $n \times \mu$, where n is the length of the homopolymeric region and μ is $1.0 \times 10^{-6}$.

[0076] Because the impact of the initial template number and the PCR efficiency on the endpoint error rate is small, it should be safe to apply the same end-point error rate estimated from the simulation experiments to molecules with different initial number and different replication efficiencies in a multiplex PCR reaction. The cutoff error rates (μ)

were empirically set as error rates at the 9999th 10000-quantiles point for each category. For two similar CDR3 sequences, A and B, of frequency NA and NB (NA >> NB) that differ in less than three positions, if NA * μ ≥ NB, where μ is the corresponding cutoff error rate, CDR3 sequence B will be excluded. Applying this filtering algorithm to the "reference filtered" sequences of template mix I, a total of 22,369 sequences, which corresponded to 281 unique CDR3 variants, were removed. For template mix II, 39,920 artifactual sequences (348 unique CDR3 variants) were identified (Table 1). Thus, the use of the PCR amplification error filter leads to a further reduction of non-authentic distinct CDR3 sequences by around 80%.

[0077] In the pool of sequences that had passed through the above filters, the inventors identified several high-abundance CDR3 variants, which differed from their most similar input template sequences at multiple positions. Because the occurrence of PCR substitution and/or indel mutation at multiple positions of CDR3 fragments is extremely rare according to simulation experiments, those CDR3 variants must arise from other source of artifacts. Intriguingly, the inventors noted that some of these sequences were composed of the fragments of two distinct input templates and exhibited clear breakpoints, which identified them as chimeras. Chimeric sequences are PCR artifacts that arise from incomplete primer extension or template switching during PCR and form mosaic-like structures. In light of this unexpected PCR artifact, the inventors developed a computational "mosaic filter." Using this filtering algorithm, the inventors identified a total of 17 and 15 chimeric sequences in template mixtures I and II, respectively. Of note, some of these CDR3 chimeras displayed sequence copy numbers >1000, indicating that the inventors' algorithm for the filter is capable of identifying high-abundance chimeric CDR3 sequences.

[0078] Application of the filtering algorithms resulted in the elimination of 99.8% of the non-authentic unique CDR3 sequences generated by high-throughput sequencing of only four defined TCR CDR3 templates. Only 62 and 73 artifactual CDR3 sequences, respectively, passed through all filters. Among these, the two most abundant CDR3

sequences were identical in both mixing experiments. Most likely they represent chimeric artifacts which escaped filtering because of a single nucleotide substitution located exactly at the breakpoint. Among the remaining erroneous CDR3, 85% (n=53) and 75% (n=55) were single reads, respectively. To eliminate this minor fraction of artifacts, the inventors propose that high-stringency data analysis of TCR immune repertoires should include an additional filter that removes single copy CDR3 reads (frequency threshold filter).

Table 1

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | | | | |
| TRAV-C | TRAV1Fo | TGCACGTACC AGACATCTGG | 12 | TGCACGTACCA GACATCTGG | 12 |
| | TRAV1Fi | AGGTCGTTTT TCTTCATTCC | 13 | GCCTCCCTCGC GCCATCAGAGG TCGTTTTTCTTC ATTCC | 14 |
| | TRAV2Fo | TCTGTAATCA CTCTGTGTCC | 15 | TCTGTAATCACT CTGTGTCC | 15 |
| | TRAV2Fi | AGGGACGATA CAACATGACC | 16 | GCCTCCCTCGC GCCATCAGAGG GACGATACAAC ATGACC | 17 |
| | TRAV3Fo | CTATTCAGTC TCTGGAAACC | 18 | CTATTCAGTCT CTGGAAACC | 18 |
| | TRAV3Fi | ATACATCACA GGGGATAACC | 19 | GCCTCCCTCGC GCCATCAGATA CATCACAGGGG ATAACC | 20 |
| | TRAV4Fo | TGTAGCCACA ACAACATTGC | 21 | TGTAGCCACAA CAACATTGC | 21 |
| | TRAV4Fi | AAAGTTACAA ACGAAGTGGC | 22 | GCCTCCCTCGC GCCATCAGAAA GTTACAAACGA AGTGGC | 23 |
| | TRAV5Fo | GCACTTACAC AGACAGCTCC | 24 | GCACTTACACA GACAGCTCC | 24 |
| | TRAV5Fi | TATGGACATG AAACAAGACC | 25 | GCCTCCCTCGC GCCATCAGTAT GGACATGAAAC AAGACC | 26 |
| | TRAV6Fo | GCAACTATAC AAACTATTCC | 27 | GCAACTATACA AACTATTCC | 27 |
| | TRAV6Fi | GTTTTCTTGC TACTCATACG | 28 | GCCTCCCTCGC GCCATCAGGTT TTCTTGCTACTC ATACG | 29 |
| | TRAV7Fo | TGCACGTACT CTGTCAGTCG | 30 | TGCACGTACTC TGTCAGTCG | 30 |
| | TRAV7Fi | GGATATGAGA AGCAGAAAGG | 31 | GCCTCCCTCGC GCCATCAGGGA TATGAGAAGCA GAAAGG | 32 |
| | TRAV8Fo | AATCTCTTCT GGTATGTSCA | 33 | AATCTCTTCTG GTATGTSCA | 33 |
| | TRAV8Fi | GGYTTTGAGG CTGAATTTA | 34 | GCCTCCCTCGC GCCATCAGGGY | 35 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | | | TTTGAGGCTGAATTTA | |
| | TRAV9Fo | GTCCAATATCCTGGAGAAGG | 36 | GTCCAATATCCTGGAGAAGG | 36 |
| | TRAV9Fi | AACCACTTCTTTCCACTTGG | 37 | GCCTCCCTCGCGCCATCAGAACCACTTCTTTCCACTTGG | 38 |
| | TRAV10Fo | AATGCAATTATACAGTGAGC | 39 | AATGCAATTATACAGTGAGC | 39 |
| | TRAV10Fi | TGAGAACACAAAGTCGAACG | 40 | GCCTCCCTCGCGCCATCAGTGAGAACACAAAGTCGAACG | 41 |
| | TRAV11Fo | TCTTAATTGTACTTATCAGG | 42 | TCTTAATTGTACTTATCAGG | 42 |
| | TRAV11Fi | TCAATCAAGCCAGAAGGAGC | 43 | GCCTCCCTCGCGCCATCAGTCAATCAAGCCAGAAGGAGC | 44 |
| | TRAV12Fo | TCAGTGTTCCAGAGGGAGCC | 45 | TCAGTGTTCCAGAGGGAGCC | 45 |
| | TRAV12Fi | ATGGAAGGTTTACAGCACAG | 46 | GCCTCCCTCGCGCCATCAGATGGAAGGTTTACAGCACAG | 47 |
| | TRAV13Fo | ACCCTGAGTGTCCAGGAGGG | 48 | ACCCTGAGTGTCCAGGAGGG | 48 |
| | TRAV13Fi | TTATAGACATTCGTTCAAAT | 49 | GCCTCCCTCGCGCCATCAGTTATAGACATTCGTTCAAAT | 50 |
| | TRAV14Fo | TGGACTGCACATATGACACC | 51 | TGGACTGCACATATGACACC | 51 |
| | TRAV14Fi | CAGCAAAATGCAACAGAAGG | 52 | GCCTCCCTCGCGCCATCAGCAGCAAAATGCAACAGAAGG | 53 |
| | TRAV16Fo | AGCTGAAGTGCAACTATTCC | 54 | AGCTGAAGTGCAACTATTCC | 54 |
| | TRAV16Fi | TCTAGAGAGAGCATCAAAGG | 55 | GCCTCCCTCGCGCCATCAGTCTAGAGAGAGCATCAAAGG | 56 |
| | TRAV17Fo | AATGCCACCATGAACTGCAG | 57 | AATGCCACCATGAACTGCAG | 57 |
| | TRAV17Fi | GAAAGAGAGAAACACAGTGG | 58 | GCCTCCCTCGCGCCATCAGGAA | 59 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | | | AGAGAGAAACA CAGTGG | |
| | TRAV18Fo | GCTCTGACAT TAAACTGCAC | 60 | GCTCTGACATT AAACTGCAC | 60 |
| | TRAV18Fi | CAGGAGACG GACAGCAGA GG | 61 | GCCTCCCTCGC GCCATCAGCAG GAGACGGACAG CAGAGG | 62 |
| | TRAV19Fo | ATGTGACCTT GGACTGTGTG | 63 | ATGTGACCTTG GACTGTGTG | 63 |
| | TRAV19Fi | GAGCAAAATG AAATAAGTGG | 64 | GCCTCCCTCGC GCCATCAGGAG CAAAATGAAAT AAGTGG | 65 |
| | TRAV20Fo | ACTGCAGTTA CACAGTCAGC | 66 | ACTGCAGTTAC ACAGTCAGC | 66 |
| | TRAV20Fi | AGAAAGAAAG GCTAAAAGCC | 67 | GCCTCCCTCGC GCCATCAGAGA AAGAAAGGCTA AAAGCC | 68 |
| | TRAV21Fo | ACTGCAGTTT CACTGATAGC | 69 | ACTGCAGTTTC ACTGATAGC | 69 |
| | TRAV21Fi | CAAGTGGAAG ACTTAATGCC | 70 | GCCTCCCTCGC GCCATCAGCAA GTGGAAGACTT AATGCC | 71 |
| | TRAV22Fo | GGGAGCCAAT TCCACGCTGC | 72 | GGGAGCCAATT CCACGCTGC | 72 |
| | TRAV22Fi | ATGGAAGATT AAGCGCCAC G | 73 | GCCTCCCTCGC GCCATCAGATG GAAGATTAAGC GCCACG | 74 |
| | TRAV23Fo | ATTTCAATTAT AAACTGTGC | 75 | ATTTCAATTATA AACTGTGC | 75 |
| | TRAV23Fi | AAGGAAGATT CACAATCTCC | 76 | GCCTCCCTCGC GCCATCAGAAG GAAGATTCACA ATCTCC | 77 |
| | TRAV24Fo | GCACCAATTT CACCTGCAGC | 78 | GCACCAATTTC ACCTGCAGC | 78 |
| | TRAV24Fi | AGGACGAATA AGTGCCACTC | 79 | GCCTCCCTCGC GCCATCAGAGG ACGAATAAGTG CCACTC | 80 |
| | TRAV25Fo | TCACCACGTA CTGCAATTCC | 81 | TCACCACGTAC TGCAATTCC | 81 |
| | TRAV25Fi | AGACTGACAT TTCAGTTTGG | 82 | GCCTCCCTCGC GCCATCAGAGA CTGACATTTCA GTTTGG | 83 |
| | TRAV26Fo | TCGACAGATT | 84 | TCGACAGATTC | 84 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | CMCTCCCAG G | | MCTCCCAGG | |
| | TRAV26Fi | GTCCAGYACC TTGATCCTGC | 85 | GCCTCCCTCGC GCCATCAGGTC CAGYACCTTGA TCCTGC | 86 |
| | TRAV27Fo | CCTCAAGTGT TTTTTCCAGC | 87 | CCTCAAGTGTT TTTTCCAGC | 87 |
| | TRAV27Fi | GTGACAGTAG TTACGGGTGG | 88 | GCCTCCCTCGC GCCATCAGGTG ACAGTAGTTAC GGGTGG | 89 |
| | TRAV29Fo | CAGCATGTTT GATTATTTCC | 90 | CAGCATGTTTG ATTATTTCC | 90 |
| | TRAV29Fi | ATCTATAAGT TCCATTAAGG | 91 | GCCTCCCTCGC GCCATCAGATC TATAAGTTCCAT TAAGG | 92 |
| | TRAV30Fo | CTCCAAGGCT TTATATTCTG | 93 | CTCCAAGGCTT TATATTCTG | 93 |
| | TRAV30Fi | ATGATATTAC TGAAGGGTG G | 94 | GCCTCCCTCGC GCCATCAGATG ATATTACTGAA GGGTGG | 95 |
| | TRAV34Fo | ACTGCACGTC ATCAAAGACG | 96 | ACTGCACGTCA TCAAAGACG | 96 |
| | TRAV34Fi | TTGATGATGC TACAGAAAGG | 97 | GCCTCCCTCGC GCCATCAGTTG ATGATGCTACA GAAAGG | 98 |
| | TRAV35Fo | TGAACTGCAC TTCTTCAAGC | 99 | TGAACTGCACT TCTTCAAGC | 99 |
| | TRAV35Fi | CTTGATAGCC TTATATAAGG | 100 | GCCTCCCTCGC GCCATCAGCTT GATAGCCTTAT ATAAGG | 101 |
| | TRAV36Fo | TCAATTGCAG TTATGAAGTG | 102 | TCAATTGCAGT TATGAAGTG | 102 |
| | TRAV36Fi | TTTATGCTAA CTTCAAGTGG | 103 | GCCTCCCTCGC GCCATCAGTTT ATGCTAACTTC AAGTGG | 104 |
| | TRAV38Fo | GCACATATGA CACCAGTGAG | 105 | GCACATATGAC ACCAGTGAG | 105 |
| | TRAV38Fi | TCGCCAAGAA GCTTATAAGC | 106 | GCCTCCCTCGC GCCATCAGTCG CCAAGAAGCTT ATAAGC | 107 |
| | TRAV39Fo | TCTACTGCAA TTATTCAACC | 108 | TCTACTGCAATT ATTCAACC | 108 |
| | TRAV39Fi | CAGGAGGGA | 109 | GCCTCCCTCGC | 110 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | CGATTAATGGC | | GCCATCAGCAGGAGGGACGATTAATGGC | |
| | TRAV40Fo | TGAACTGCACATACACATCC | 111 | TGAACTGCACATACACATCC | 111 |
| | TRAV40Fi | ACAGCAAAAACTTCGGAGGC | 112 | GCCTCCCTCGCGCCATCAGACAGCAAAAACTTCGGAGGC | 113 |
| | TRAV41Fo | AACTGCAGTTACTCGGTAGG | 114 | AACTGCAGTTACTCGGTAGG | 114 |
| | TRAV41Fi | AAGCATGGAAGATTAATTGC | 115 | GCCTCCCTCGCGCCATCAGAAGCATGGAAGATTAATTGC | 116 |
| | | | | | |
| | TRACRo | GCAGACAGACTTGTCACTGG | 117 | GCAGACAGACTTGTCACTGG | 117 |
| | TRACRi | AGTCTCTCAGCTGGTACACG | 118 | GCCTTGCCAGCCCGCTCAGAGTCTCTCAGCTGGTACACG | 119 |
| | | | | | |
| TRBV-C | TRBV1Fo | AATGAAACGTGAGCATCTGG | 120 | AATGAAACGTGAGCATCTGG | 120 |
| | TRBV1Fi | CATTGAAAACAAGACTGTGC | 121 | GCCTCCCTCGCGCCATCAGCATTGAAAACAAGACTGTGC | 122 |
| | TRBV2Fo | GTGTCCCCATCTCTAATCAC | 123 | GTGTCCCCATCTCTAATCAC | 123 |
| | TRBV2Fi | TGAAATCTCAGAGAAGTCTG | 124 | GCCTCCCTCGCGCCATCAGTGAAATCTCAGAGAAGTCTG | 125 |
| | TRBV3Fo | TATGTATTGGTATAAACAGG | 126 | TATGTATTGGTATAAACAGG | 126 |
| | TRBV3Fi | CTCTAAGAAATTTCTGAAGA | 127 | GCCTCCCTCGCGCCATCAGCTCTAAGAAATTTCTGAAGA | 128 |
| | TRBV4Fo | GTCTTTGAAATGTGAACAAC | 129 | GTCTTTGAAATGTGAACAAC | 129 |
| | TRBV4Fi | GGAGCTCATGTTTGTCTACA | 130 | GCCTCCCTCGCGCCATCAGGGAGCTCATGTTTGTCTACA | 131 |
| | TRBV5Fo | GATCAAAACGAGAGGACAGC | 132 | GATCAAAACGAGAGGACAGC | 132 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | TRBV5aFi | CAGGGGCCCCAGTTTATCTT | 133 | GCCTCCCTCGCGCCATCAGCAGGGGCCCCAGTTTATCTT | 134 |
| | TRBV5bFi | GAAACARAGGAAACTTCCCT | 135 | GCCTCCCTCGCGCCATCAGGAAACARAGGAAACTTCCCT | 136 |
| | TRBV6aFo | GTGTGCCCAGGATATGAACC | 137 | GTGTGCCCAGGATATGAACC | 137 |
| | TRBV6bFo | CAGGATATGAGACATAATGC | 138 | CAGGATATGAGACATAATGC | 138 |
| | TRBV6aFi | GGTATCGACAAGACCCAGGC | 139 | GCCTCCCTCGCGCCATCAGGGTATCGACAAGACCCAGGC | 140 |
| | TRBV6bFi | TAGACAAGATCTAGGACTGG | 141 | GCCTCCCTCGCGCCATCAGTAGACAAGATCTAGGACTGG | 142 |
| | TRBV7Fo | CTCAGGTGTGATCCAATTTC | 143 | CTCAGGTGTGATCCAATTTC | 143 |
| | TRBV7aFi | TCTAATTTACTTCCAAGGCA | 144 | GCCTCCCTCGCGCCATCAGTCTAATTTACTTCCAAGGCA | 145 |
| | TRBV7bFi | TCCCAGAGTGATGCTCAACG | 146 | GCCTCCCTCGCGCCATCAGTCCCAGAGTGATGCTCAACG | 147 |
| | TRBV7cFi | ACTTACTTCAATTATGAAGC | 148 | GCCTCCCTCGCGCCATCAGACTTACTTCAATTATGAAGC | 149 |
| | TRBV7dFi | CCAGAATGAAGCTCAACTAG | 150 | GCCTCCCTCGCGCCATCAGCCAGAATGAAGCTCAACTAG | 151 |
| | TRBV9Fo | GAGACCTCTCTGTGTACTGG | 152 | GAGACCTCTCTGTGTACTGG | 152 |
| | TRBV9Fi | CTCATTCAGTATTATAATGG | 153 | GCCTCCCTCGCGCCATCAGCTCATTCAGTATTATAATGG | 154 |
| | TRBV10Fo | GGAATCACCCAGAGCCCAAG | 155 | GGAATCACCCAGAGCCCAAG | 155 |
| | TRBV10Fi | GACATGGGCTGAGGCTGATC | 156 | GCCTCCCTCGCGCCATCAGGACATGGGCTGAGGCTGATC | 157 |
| | TRBV11Fo | CCTAAGGATC | 158 | CCTAAGGATCG | 158 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | GATTTTCTGC | | ATTTTCTGC | |
| | TRBV11Fi | ACTCTCAAGA TCCAGCCTGC | 159 | GCCTCCCTCGC GCCATCAGACT CTCAAGATCCA GCCTGC | 160 |
| | TRBV12Fo | AGGTGACAGA GATGGGACAA | 161 | AGGTGACAGAG ATGGGACAA | 161 |
| | TRBV12aFi | TGCAGGGACT GGAATTGCTG | 162 | GCCTCCCTCGC GCCATCAGTGC AGGGACTGGAA TTGCTG | 163 |
| | TRBV12bFi | GTACAGACAG ACCATGATGC | 164 | GCCTCCCTCGC GCCATCAGGTA CAGACAGACCA TGATGC | 165 |
| | TRBV13Fo | CTATCCTATC CCTAGACACG | 166 | CTATCCTATCC CTAGACACG | 166 |
| | TRBV13Fi | AAGATGCAGA GCGATAAAGG | 167 | GCCTCCCTCGC GCCATCAGAAG ATGCAGAGCGA TAAAGG | 168 |
| | TRBV14Fo | AGATGTGACC CAATTTCTGG | 169 | AGATGTGACCC AATTTCTGG | 169 |
| | TRBV14Fi | AGTCTAAACA GGATGAGTCC | 170 | GCCTCCCTCGC GCCATCAGAGT CTAAACAGGAT GAGTCC | 171 |
| | TRBV15Fo | TCAGACTTTG AACCATAACG | 172 | TCAGACTTTGA ACCATAACG | 172 |
| | TRBV15Fi | AAAGATTTTA ACAATGAAGC | 173 | GCCTCCCTCGC GCCATCAGAAA GATTTTAACAAT GAAGC | 174 |
| | TRBV16Fo | TATTGTGCCC CAATAAAAGG | 175 | TATTGTGCCCC AATAAAAGG | 175 |
| | TRBV16Fi | AATGTCTTTG ATGAAACAGG | 176 | GCCTCCCTCGC GCCATCAGAAT GTCTTTGATGA AACAGG | 177 |
| | TRBV17Fo | ATCCATCTTC TGGTCACATG | 178 | ATCCATCTTCT GGTCACATG | 178 |
| | TRBV17Fi | AACATTGCAG TTGATTCAGG | 179 | GCCTCCCTCGC GCCATCAGAAC ATTGCAGTTGA TTCAGG | 180 |
| | TRBV18Fo | GCAGCCCAAT GAAAGGACAC | 181 | GCAGCCCAATG AAAGGACAC | 181 |
| | TRBV18Fi | AATATCATAG ATGAGTCAGG | 182 | GCCTCCCTCGC GCCATCAGAAT ATCATAGATGA GTCAGG | 183 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | TRBV19Fo | TGAACAGAATTTGAACCACG | 184 | TGAACAGAATTTGAACCACG | 184 |
| | TRBV19Fi | TTTCAGAAAGGAGATATAGC | 185 | GCCTCCCTCGCGCCATCAGTTTCAGAAAGGAGATATAGC | 186 |
| | TRBV20Fo | TCGAGTGCCGTTCCCTGGAC | 187 | TCGAGTGCCGTTCCCTGGAC | 187 |
| | TRBV20Fi | GATGGCAACTTCCAATGAGG | 188 | GCCTCCCTCGCGCCATCAGGATGGCAACTTCCAATGAGG | 189 |
| | TRBV21Fo | GCAAAGATGGATTGTGTTCC | 190 | GCAAAGATGGATTGTGTTCC | 190 |
| | TRBV21Fi | CGCTGGAAGAAGAGCTCAAG | 191 | GCCTCCCTCGCGCCATCAGCGCTGGAAGAAGAGCTCAAG | 192 |
| | TRBV23Fo | CATTTGGTCAAAGGAAAAGG | 193 | CATTTGGTCAAAGGAAAAGG | 193 |
| | TRBV23Fi | GAATGAACAAGTTCTTCAAG | 194 | GCCTCCCTCGCGCCATCAGGAATGAACAAGTTCTTCAAG | 195 |
| | TRBV24Fo | ATGCTGGAATGTTCTCAGAC | 196 | ATGCTGGAATGTTCTCAGAC | 196 |
| | TRBV24Fi | GTCAAAGATATAAACAAAGG | 197 | GCCTCCCTCGCGCCATCAGGTCAAAGATATAAACAAAGG | 198 |
| | TRBV25Fo | CTCTGGAATGTTCTCAAACC | 199 | CTCTGGAATGTTCTCAAACC | 199 |
| | TRBV25Fi | TAATTCCACAGAGAAGGGAG | 200 | GCCTCCCTCGCGCCATCAGTAATTCCACAGAGAAGGGAG | 201 |
| | TRBV26Fo | CCCAGAATATGAATCATGTT | 202 | CCCAGAATATGAATCATGTT | 202 |
| | TRBV26Fi | ATTCACCTGGCACTGGGAGC | 203 | GCCTCCCTCGCGCCATCAGATTCACCTGGCACTGGGAGC | 204 |
| | TRBV27Fo | TTGTTCTCAGAATATGAACC | 205 | TTGTTCTCAGAATATGAACC | 205 |
| | TRBV27Fi | TGAGGTGACTGATAAGGGAG | 206 | GCCTCCCTCGCGCCATCAGTGAGGTGACTGATAAGGGAG | 207 |
| | TRBV28Fo | ATGTGTCCAGGATATGGACC | 208 | ATGTGTCCAGGATATGGACC | 208 |
| | TRBV28Fi | AAAAGGAGAT | 209 | GCCTCCCTCGC | 210 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| | | ATTCCTGAGG | | GCCATCAGAAA AGGAGATATTC CTGAGG | |
| | TRBV29Fo | TCACCATGAT GTTCTGGTAC | 211 | TCACCATGATG TTCTGGTAC | 211 |
| | TRBV29Fi | CTGGACAGAG CCTGACACTG | 212 | GCCTCCCTCGC GCCATCAGCTG GACAGAGCCTG ACACTG | 213 |
| | TRBV30Fo | TGTGGAGGG AACATCAAAC C | 214 | TGTGGAGGGAA CATCAAACC | 214 |
| | TRBV30Fi | TTCTACTCCG TTGGTATTGG | 215 | GCCTCCCTCGC GCCATCAGTTC TACTCCGTTGG TATTGG | 216 |
| | | | | | |
| | TRBCRo | GTGTGGCCTT TTGGGTGTGG | 217 | GTGTGGCCTTT TGGGTGTGG | 217 |
| | TRBCRi | TCTGATGGCT CAAACACAGC | 218 | GCCTTGCCAGC CCGCTCAGTCT GATGGCTCAAA CACAGC | 219 |
| | | | | | |
| TRDV-C | TRDV1Fo | TGTATGAAAC AAGTTGGTGG | 220 | TGTATGAAACA AGTTGGTGG | 220 |
| | TRDV1Fi | CAGAATGCAA AAAGTGGTCG | 221 | GCCTCCCTCGC GCCATCAGCAG AATGCAAAAAG TGGTCG | 222 |
| | TRDV2Fo | ATGAAAGGAG AAGCGATCGG | 223 | ATGAAAGGAGA AGCGATCGG | 223 |
| | TRDV2Fi | TGGTTTCAAA GACAATTTCC | 224 | GCCTCCCTCGC GCCATCAGTGG TTTCAAAGACA ATTTCC | 225 |
| | TRDV3Fo | GACACTGTAT ATTCAAATCC | 226 | GACACTGTATA TTCAAATCC | 226 |
| | TRDV3Fi | GCAGATTTTA CTCAAGGACG | 227 | GCCTCCCTCGC GCCATCAGGCA GATTTTACTCAA GGACG | 228 |
| | | | | | |
| | TRDCRo | AGACAAGCGA CATTTGTTCC | 229 | AGACAAGCGAC ATTTGTTCC | 229 |
| | TRDCRi | ACGGATGGTT TGGTATGAGG | 230 | GCCTTGCCAGC CCGCTCAGACG GATGGTTTGGT ATGAGG | 231 |
| | | | | | |

| Locus | Primer Name | Sequence | SEQ ID NO. | Sequence | SEQ ID NO. |
|---|---|---|---|---|---|
| TRGV-C | TRGV1-5Fo | GGGTCATCTGCTGAAATCAC | 232 | GGGTCATCTGCTGAAATCAC | 232 |
| | TRGV1-5,8Fi | AGGAGGGGAAGGCCCCACAG | 233 | GCCTCCCTCGCGCCATCAGAGGAGGGGAAGGCCCCACAG | 234 |
| | TRGV8Fo | GGGTCATCAGCTGTAATCAC | 235 | GGGTCATCAGCTGTAATCAC | 235 |
| | TRGV5pFi | AGGAGGGGAAGACCCCACAG | 236 | GCCTCCCTCGCGCCATCAGAGGAGGGGAAGACCCCACAG | 237 |
| | TRGV9Fo | AGCCCGCCTGGAATGTGTGG | 238 | AGCCCGCCTGGAATGTGTGG | 238 |
| | TRGV9Fi | GCACTGTCAGAAAGGAATCC | 239 | GCCTCCCTCGCGCCATCAGGCACTGTCAGAAAGGAATCC | 240 |
| | TRGV10Fo | AAGAAAAGTATTGACATACC | 241 | AAGAAAAGTATTGACATACC | 241 |
| | TRGV10Fi | ATATTGTCTCAACAAAATCC | 242 | GCCTCCCTCGCGCCATCAGATATTGTCTCAACAAAATCC | 243 |
| | TRGV11Fo | AGAGTGCCCACATATCTTGG | 244 | AGAGTGCCCACATATCTTGG | 244 |
| | TRGV11Fi | GCTCAAGATTGCTCAGGTGG | 245 | GCCTCCCTCGCGCCATCAGGCTCAAGATTGCTCAGGTGG | 246 |
| | | | | | |
| | TRGCRo | GGATCCCAGAATCGTGTTGC | 247 | GGATCCCAGAATCGTGTTGC | 247 |
| | TRGCRi | GGTATGTTCCAGCCTTCTGG | 248 | GCCTTGCCAGCCCGCTCAGGGTATGTTCCAGCCTTCTGG | 249 |

**Table 2**

| Locus | Primer Name | Sequence | SEQ ID NO. | Ordered | SEQ ID NO. |
|---|---|---|---|---|---|
| | | | | | |
| IgHV-J | IgHV1aFo | AGTGAAGGTCTC CTGCAAGG | 250 | AGTGAAGGTCTC CTGCAAGG | 250 |
| | IgHV1bFo | AGTGAAGGTTTC CTGCAAGG | 251 | AGTGAAGGTTTC CTGCAAGG | 251 |
| | IgHV1aFi | AGTTCCAGGGCA GAGTCAC | 252 | GCCTCCCTCGCG CCATCAGAGTTC CAGGGCAGAGTC AC | 253 |
| | IgHV1bFi | AGTTTCAGGGCA GGGTCAC | 254 | GCCTCCCTCGCG CCATCAGAGTTT CAGGGCAGGGTC AC | 255 |
| | IgHV1cFi | AGTTCCAGGAAA GAGTCAC | 256 | GCCTCCCTCGCG CCATCAGAGTTC CAGGAAAGAGTC AC | 257 |
| | IgHV1dFi | AATTCCAGGACA GAGTCAC | 258 | GCCTCCCTCGCG CCATCAGAATTC CAGGACAGAGTC AC | 259 |
| | IgHV2Fo | TCTCTGGGTTCT CACTCAGC | 260 | TCTCTGGGTTCT CACTCAGC | 260 |
| | IgHV2Fi | AAGGCCCTGGAG TGGCTTGC | 261 | GCCTCCCTCGCG CCATCAGAAGGC CCTGGAGTGGCT TGC | 262 |
| | IgHV3aFo | TCCCTGAGACTC TCCTGTGC | 263 | TCCCTGAGACTC TCCTGTGC | 263 |
| | IgHV3bFo | CTCTCCTGTGCA GCCTCTGG | 264 | CTCTCCTGTGCA GCCTCTGG | 264 |
| | IgHV3cFo | GGTCCCTGAGAC TCTCCTGT | 265 | GGTCCCTGAGAC TCTCCTGT | 265 |
| | IgHV3dFo | CTGAGACTCTCC TGTGTAGC | 266 | CTGAGACTCTCC TGTGTAGC | 266 |
| | IgHV3aFi | CTCCAGGGAAGG GGCTGG | 267 | GCCTCCCTCGCG CCATCAGCTCCA GGGAAGGGGCT GG | 268 |
| | IgHV3bFi | GGCTCCAGGCAA GGGGCT | 269 | GCCTCCCTCGCG CCATCAGGGCTC CAGGCAAGGGGC T | 270 |
| | IgHV3cFi | ACTGGGTCCGCC AGGCTCC | 271 | GCCTCCCTCGCG CCATCAGACTGG GTCCGCCAGGCT CC | 272 |
| | IgHV3dFi | GAAGGGGCTGGA GTGGGT | 273 | GCCTCCCTCGCG CCATCAGGAAGG GGCTGGAGTGGG T | 274 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Ordered | SEQ ID NO. |
|---|---|---|---|---|---|
| | IgHV3eFi | AAAAGGTCTGGAGTGGGT | 275 | GCCTCCCTCGCGCCATCAGAAAAGGTCTGGAGTGGGT | 276 |
| | IgHV4Fo | AGACCCTGTCCCTCACCTGC | 277 | AGACCCTGTCCCTCACCTGC | 277 |
| | IgHV4Fi | AGGGVCTGGAGTGGATTGGG | 278 | GCCTCCCTCGCGCCATCAGAGGGVCTGGAGTGGATTGGG | 279 |
| | IgHV5Fo | GCGCCAGATGCCCGGGAAAG | 280 | GCGCCAGATGCCCGGGAAAG | 280 |
| | IgHV5Fi | GGCCASGTCACCATCTCAGC | 281 | GCCTCCCTCGCGCCATCAGGGCCASGTCACCATCTCAGC | 282 |
| | IgHV6Fo | CCGGGGACAGTGTCTCTAGC | 283 | CCGGGGACAGTGTCTCTAGC | 283 |
| | IgHV6Fi | GCCTTGAGTGGCTGGGAAGG | 284 | GCCTCCCTCGCGCCATCAGGCCTTGAGTGGCTGGGAAGG | 285 |
| | IgHV7Fo | GTTTCCTGCAAGGCTTCTGG | 286 | GTTTCCTGCAAGGCTTCTGG | 286 |
| | IgHV7Fi | GGCTTGAGTGGATGGGATGG | 287 | GCCTCCCTCGCGCCATCAGGGCTTGAGTGGATGGGATGG | 288 |
| | | | | | |
| | IgHJRo | ACCTGAGGAGACGGTGACC | 289 | ACCTGAGGAGACGGTGACC | 289 |
| | IgHJ1Ri | CAGTGCTGGAAGTATTCAGC | 290 | GCCTTGCCAGCCCGCTCAGCAGTGCTGGAAGTATTCAGC | 291 |
| | IgHJ2Ri | AGAGATCGAAGTACCAGTAG | 292 | GCCTTGCCAGCCCGCTCAGAGAGATCGAAGTACCAGTAG | 293 |
| | IgHJ3Ri | CCCCAGATATCAAAAGCATC | 294 | GCCTTGCCAGCCCGCTCAGCCCCAGATATCAAAAGCATC | 295 |
| | IgHJ4Ri | GGCCCCAGTAGTCAAAGTAG | 296 | GCCTTGCCAGCCCGCTCAGGGCCCCAGTAGTCAAAGTAG | 297 |
| | IgHJ5Ri | CCCAGGGGTCGAACCAGTTG | 298 | GCCTTGCCAGCCCGCTCAGCCCAGGGGTCGAACCAGTTG | 299 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Ordered | SEQ ID NO. |
|---|---|---|---|---|---|
| | IgHJ6Ri | CCCAGACGTCCATGTAGTAG | 300 | GCCTTGCCAGCCCGCTCAGCCCAGACGTCCATGTAGTAG | 301 |
| | | | | | |
| **IgKV-C** | IgKV1Fo | TAGGAGACAGAGTCACCATC | 302 | TAGGAGACAGAGTCACCATC | 302 |
| | IgKV1Fi | TTCAGYGRCAGTGGATCTGG | 303 | GCCTCCCTCGCGCCATCAGTTCAGYGRCAGTGGATCTGG | 304 |
| | IgKV2Fo | GGAGAGCCGGCCTCCATCTC | 305 | GGAGAGCCGGCCTCCATCTC | 305 |
| | IgKV2aFi | TGGTACCTGCAGAAGCCAGG | 306 | GCCTCCCTCGCGCCATCAGTGGTACCTGCAGAAGCCAGG | 307 |
| | IgKV2bFi | CTTCAGCAGAGGCCAGGCCA | 308 | GCCTCCCTCGCGCCATCAGCTTCAGCAGAGGCCAGGCCA | 309 |
| | IgKV3-7Fo | GCCTGGTACCAGCAGAAACC | 310 | GCCTGGTACCAGCAGAAACC | 310 |
| | IgKV3Fi | GCCAGGTTCAGTGGCAGTGG | 311 | GCCTCCCTCGCGCCATCAGGCCAGGTTCAGTGGCAGTGG | 312 |
| | IgKV6-7Fi | TCGAGGTTCAGTGGCAGTGG | 313 | GCCTCCCTCGCGCCATCAGTCGAGGTTCAGTGGCAGTGG | 314 |
| | IgKV4-5Fi | GACCGATTCAGTGGCAGCGG | 315 | GCCTCCCTCGCGCCATCAGGACCGATTCAGTGGCAGCGG | 316 |
| | | | | | |
| | IgKCRo | TTCAACTGCTCATCAGATGG | 317 | TTCAACTGCTCATCAGATGG | 317 |
| | IgKCRi | ATGAAGACAGATGGTGCAGC | 318 | GCCTTGCCAGCCCGCTCAGATGAAGACAGATGGTGCAGC | 319 |
| | | | | | |
| **IgLV-C** | IgLV1aFo | GGGCAGAGGGTCACCATCTC | 320 | GGGCAGAGGGTCACCATCTC | 320 |
| | IgLV1bFo | GGACAGAAGGTCACCATCTC | 321 | GGACAGAAGGTCACCATCTC | 321 |
| | IgLV1aFi | TGGTACCAGCAGCTCCCAGG | 322 | GCCTCCCTCGCGCCATCAGTGGTACCAGCAGCTCCCAGG | 323 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Ordered | SEQ ID NO. |
|---|---|---|---|---|---|
| | IgLV1bFi | TGGTACCAGCAG CTTCCAGG | 324 | GCCTCCCTCGCG CCATCAGTGGTA CCAGCAGCTTCC AGG | 325 |
| | IgLV2Fo | CTGCACTGGAAC CAGCAGTG | 326 | CTGCACTGGAAC CAGCAGTG | 326 |
| | IgLV2Fi | TCTCTGGCTCCA AGTCTGGC | 327 | GCCTCCCTCGCG CCATCAGTCTCT GGCTCCAAGTCT GGC | 328 |
| | IgLV3aFo | ACCAGCAGAAGC CAGGCCAG | 329 | ACCAGCAGAAGC CAGGCCAG | 329 |
| | IgLV3bFo | GAAGCCAGGACA GGCCCCTG | 330 | GAAGCCAGGACA GGCCCCTG | 330 |
| | IgLV3aFi | CTGAGCGATTCT CTGGCTCC | 331 | GCCTCCCTCGCG CCATCAGCTGAG CGATTCTCTGGC TCC | 332 |
| | IgLV3bFi | TTCTCTGGGTCC ACCTCAGG | 333 | GCCTCCCTCGCG CCATCAGTTCTCT GGGTCCACCTCA GG | 334 |
| | IgLV3cFi | TTCTCTGGCTCC AGCTCAGG | 335 | GCCTCCCTCGCG CCATCAGTTCTCT GGCTCCAGCTCA GG | 336 |
| | IgLV4Fo | TCGGTCAAGCTC ACCTGCAC | 337 | TCGGTCAAGCTC ACCTGCAC | 337 |
| | IgLV4Fi | GGGCTGACCGCT ACCTCACC | 358 | GCCTCCCTCGCG CCATCAGGGGCT GACCGCTACCTC ACC | 338 |
| | IgLV5Fo | CAGCCTGTGCTG ACTCAGCC | 339 | CAGCCTGTGCTG ACTCAGCC | 339 |
| | IgLV5Fi | CCAGCCGCTTCT CTGGATCC | 340 | GCCTCCCTCGCG CCATCAGCCAGC CGCTTCTCTGGA TCC | 341 |
| | IgLV6Fo | CCATCTCCTGCA CCCGCAGC | 342 | CCATCTCCTGCA CCCGCAGC | 342 |
| | IgLV7-8Fo | TCCCCWGGAGG GACAGTCAC | 343 | TCCCCWGGAGG GACAGTCAC | 343 |
| | IgLV9,11Fo | CTCMCCTGCACC CTGAGCAG | 344 | CTCMCCTGCACC CTGAGCAG | 344 |
| | IgLV10Fo | AGACCGCCACAC TCACCTGC | 345 | AGACCGCCACAC TCACCTGC | 345 |
| | IgLV6,8Fi | CTGATCGSTTCTC TGGCTCC | 346 | GCCTCCCTCGCG CCATCAGCTGAT CGSTTCTCTGGC TCC | 347 |
| | IgLV7Fi | CTGCCCGGTTCT | 348 | CTGCCCGGTTCT | 348 |

| Locus | Primer Name | Sequence | SEQ ID NO. | Ordered | SEQ ID NO. |
|---|---|---|---|---|---|
| | | CAGGCTCC | | CAGGCTCC | |
| | IgLV9Fi | ATCCAGGAAGAG GATGAGAG | 349 | GCCTCCCTCGCG CCATCAGATCCA GGAAGAGGATGA GAG | 350 |
| | IgLV10-11Fi | CTCCAGCCTGAG GACGAGGC | 351 | GCCTCCCTCGCG CCATCAGCTCCA GCCTGAGGACGA GGC | 352 |
| | | | | | |
| | IgLC1-7Ro | GCTCCCGGGTAG AAGTCACT | 353 | GCTCCCGGGTAG AAGTCACT | 353 |
| | IgLC1-7Ri | AGTGTGGCCTTG TTGGCTTG | 354 | GCCTTGCCAGCC CGCTCAGAGTGT GGCCTTGTTGGC TTG | 355 |
| | | | | | |
| | 454A | GCCTCCCTCGCG CCATCAG | 356 | GCCTCCCTCGCG CCATCAG | 356 |
| | | | | | |
| | 454B | GCCTTGCCAGCC CGCTCAG | 357 | GCCTTGCCAGCC CGCTCAG | 357 |

## CLAIMS

Now, therefore, the following is claimed:

1. A method for evaluating changes in immune response cell populations and associating those changes with a specific disease, the method comprising the steps of:

(a) isolating a subpopulation of white blood cells from at least one human or animal subject;

(b) isolating RNA from the subpopulation of cells;

(c) amplifying the RNA using RT-PCR in a first amplification reaction to produce amplicons using nested primers, at least a portion of the nested primers comprising additional nucleotides to incorporate into a resulting amplicon a binding site for a communal primer;

(d) separating the amplicons from the first amplification reaction from one or more unused primers from the first amplification reaction;

(e) amplifying, by the addition of communal primers in a second amplification reaction, the amplicons of the first amplification reaction having at least one binding site for a communal primer; and

(f) sequencing the amplicons of the second amplification reaction to identify antibody and/or receptor rearrangements in the subpopulation of cells.

2. The method of claim 1, wherein the product of the second amplification reaction is a polynucleotide comprising the complementarity determining region 3 (CDR3).

3. The method of claim 1, wherein the step of isolating a subpopulation of white blood cells is performed by flow cytometry.

4. The method of claim 1, wherein the subpopulation of white blood cells comprises T cells.

5. The method of claim 4, wherein the T cells are selected from the group consisting of naïve T cells, mature T cells and memory T cells.

6. The method of claim 1, wherein the subpopulation of white blood cells comprises B cells.

7. The method of claim 6, wherein the B cells are selected from the group consisting of naïve B cells, mature B cells and memory B cells.

8. The method of claim 1, wherein the rearrangements in the subpopulations of cells are selected from the group consisting of rearrangements of B-cell immunoglobulin heavy chain (IgH), B-cell kappa, B-cell lambda light chains, T-cell receptor Beta, T-cell Gamma and T-cell Delta.

9. The method of claim 1, further comprising the steps of:

(g) comparing the rearrangements identified for a population of individuals to whom a vaccine has been administered with the rearrangements identified for a population of individuals to whom the vaccine was not administered; and

(h) evaluating the efficacy of the vaccine in producing an immune response.

10. The method of claim 1, further comprising the steps of:

(g) comparing the rearrangements identified for a population of normal individuals with the rearrangements identified for a population of individuals who have been diagnosed with a disease;

(h) determining if there is a correlation between a specific rearrangement or set of rearrangements and the disease.

11. A method for analyzing semi-quantitative sequence information to provide one or more immune status reports for a human or animal, the method comprising the steps of:

(a) identifying one or more distinct CDR3 sequences that are shared between a subject's immunoprofile and a cumulative immunoprofile from a disease library stored in a database;

(b) summing a total number of a subject's detected sequences corresponding to those shared distinct CDR3 sequences;

(c) computing the percentage of the total number of detected sequences in the subject's immunoprofile that are representative of those distinct CDR3s shared between the subject's immunoprofile and the disease library to create one or more original sharing indices;

(d) randomly selecting sequences from a public library stored in a database to form a sub-library, the sub-library comprising a number of distinct CDR3 sequences that is approximately equal to the number of distinct CDR3 sequences in the disease library;

(e) identifying one or more distinct CDR3 sequences that are shared between the subject's immunoprofile and the sub-library;

(f) summing a total number of detected sequences corresponding to those shared CDR3 sequences and calculating a percentage of the total number of detected sequences in the subject's immunoprofile that are shared between the subject's immunoprofile and the sub-library to create a sampling sharing index;

(g) repeating steps (d)-(f) at least 1000 or more times; and

(h) estimating the P-value as the fraction of times the sampling sharing indices are greater than or equal to the original sharing index between a patient's immunoprofile and a disease library.

12. A method for developing a database of personal immunorepertoires, the method comprising the steps of:

(a) amplifying and sequencing one or more RNAs from a subpopulation of white blood cells from one or more individuals;

(b) inputting the sequences into a database to provide data which may be stored on a computer, server, or other electronic storage device;

(c) inputting identifying information and characteristics for an individual corresponding to the sequences of the one or more RNAs as data which may also be stored on a computer, server, or other electronic storage device, and

(d) evaluating the data of step (b) and step (e) for one or more individuals to determine whether a correlation exists between the one or more RNA sequences and one or more characteristics of the individual corresponding to the sequence(s).

13. The method of claim 12, wherein the identifying information is selected from the group consisting of a patient identification number, a code comprising the patient's HLA type, a disease code comprising one or more clinical diagnoses that may have been made, a "staging code" comprising the date of the sample, a cell type code comprising the type of cell subpopulation from which the RNA was amplified and sequenced, and one or more sequence codes comprising the sequences identified for the sample.

14. The method of claim 12, wherein the subpopulation of white blood cells comprises T cells.

15. The method of claim 14, wherein the T cells are selected from the group consisting of naïve T cells, mature T cells and memory T cells.

16. The method of claim 12, wherein the subpopulation of white blood cells comprises B cells.

17. The method of claim 16, wherein the B cells are selected from the group consisting of naïve B cells, mature B cells and memory B cells.

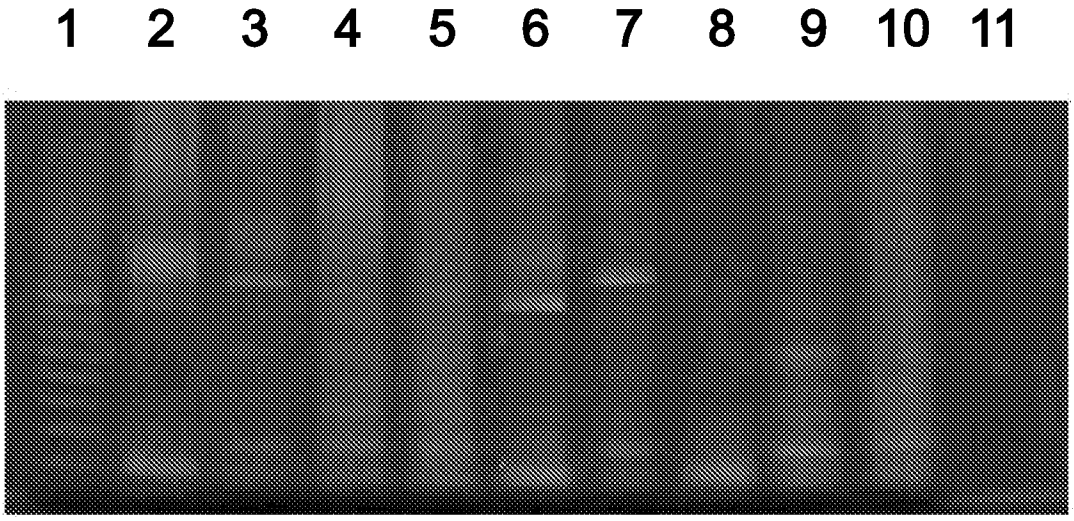



*FIG. 1a*

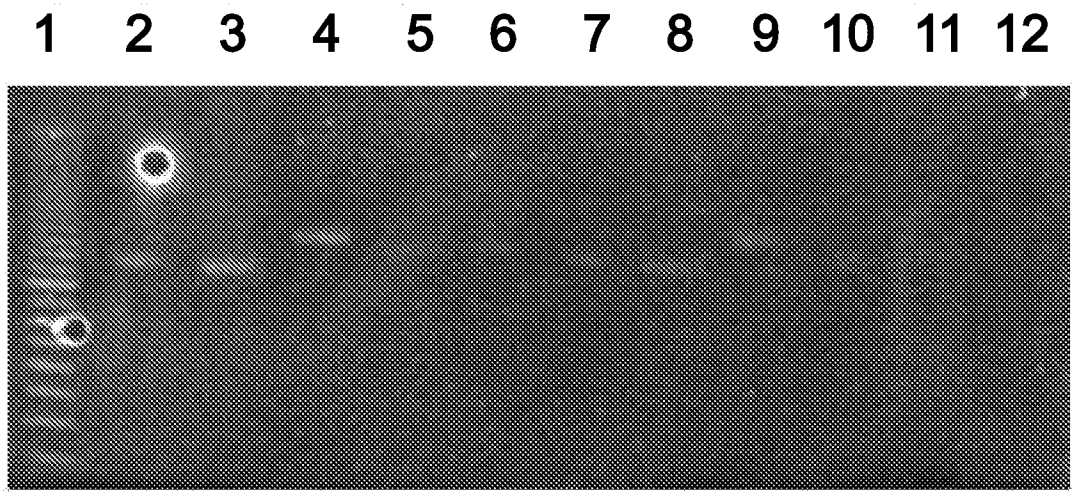



*FIG. 1b*

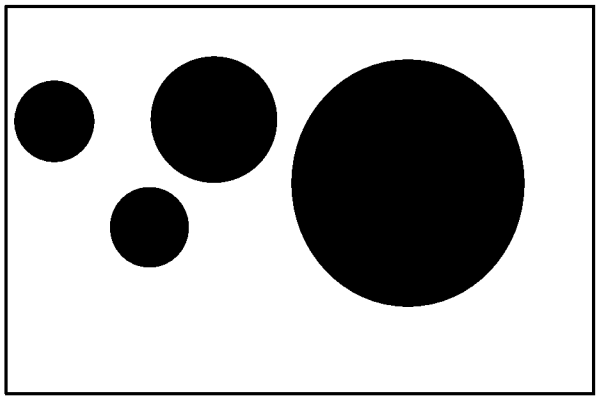


FIG. 2a

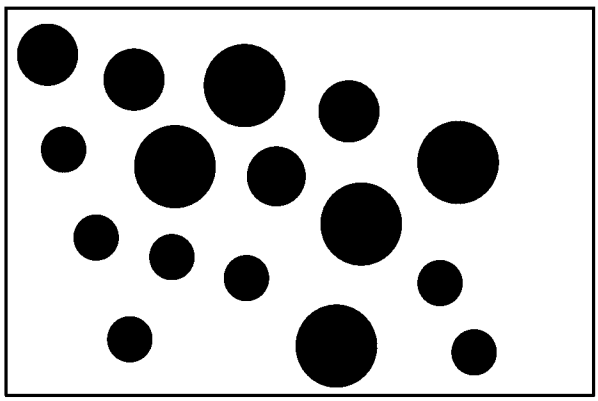


FIG. 2b

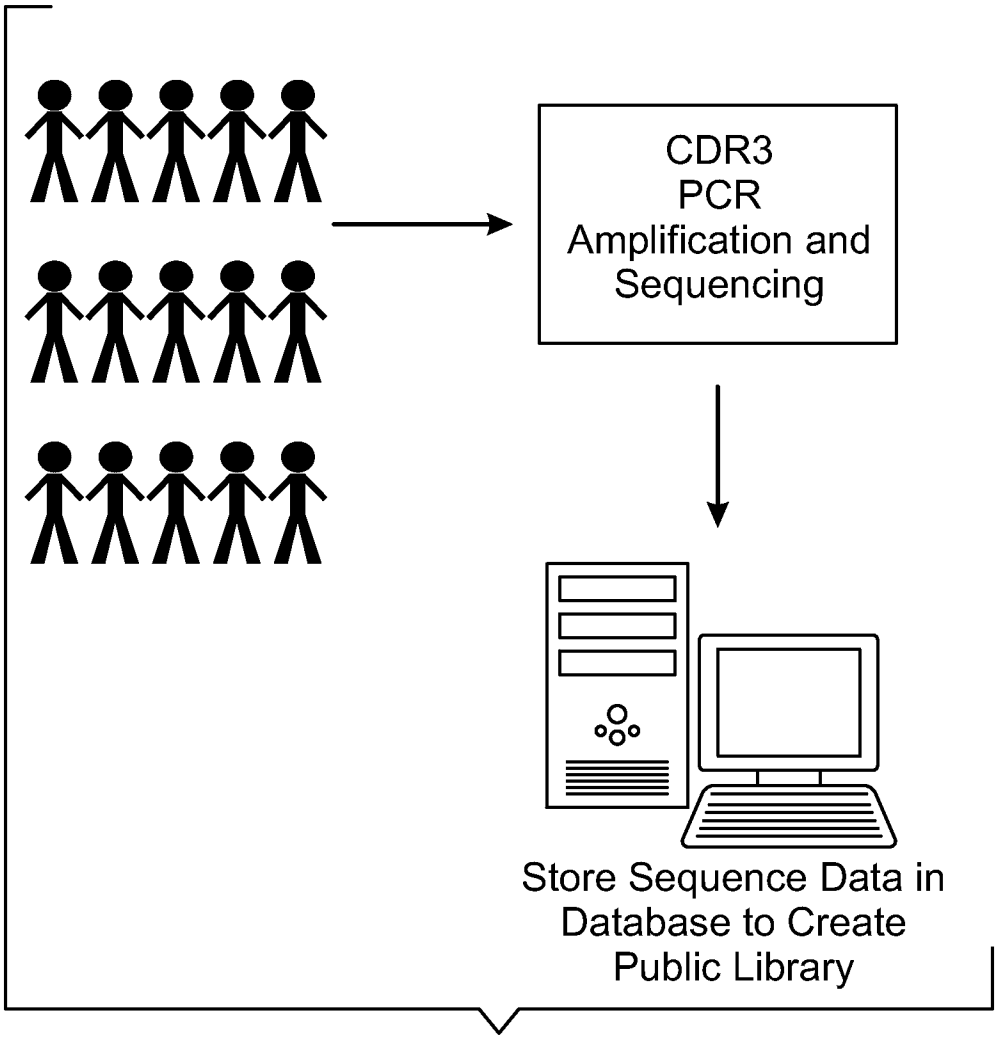

CDR3
PCR
Amplification and
Sequencing
Store Sequence Data in
Database to Create
Public Library

*FIG. 3*

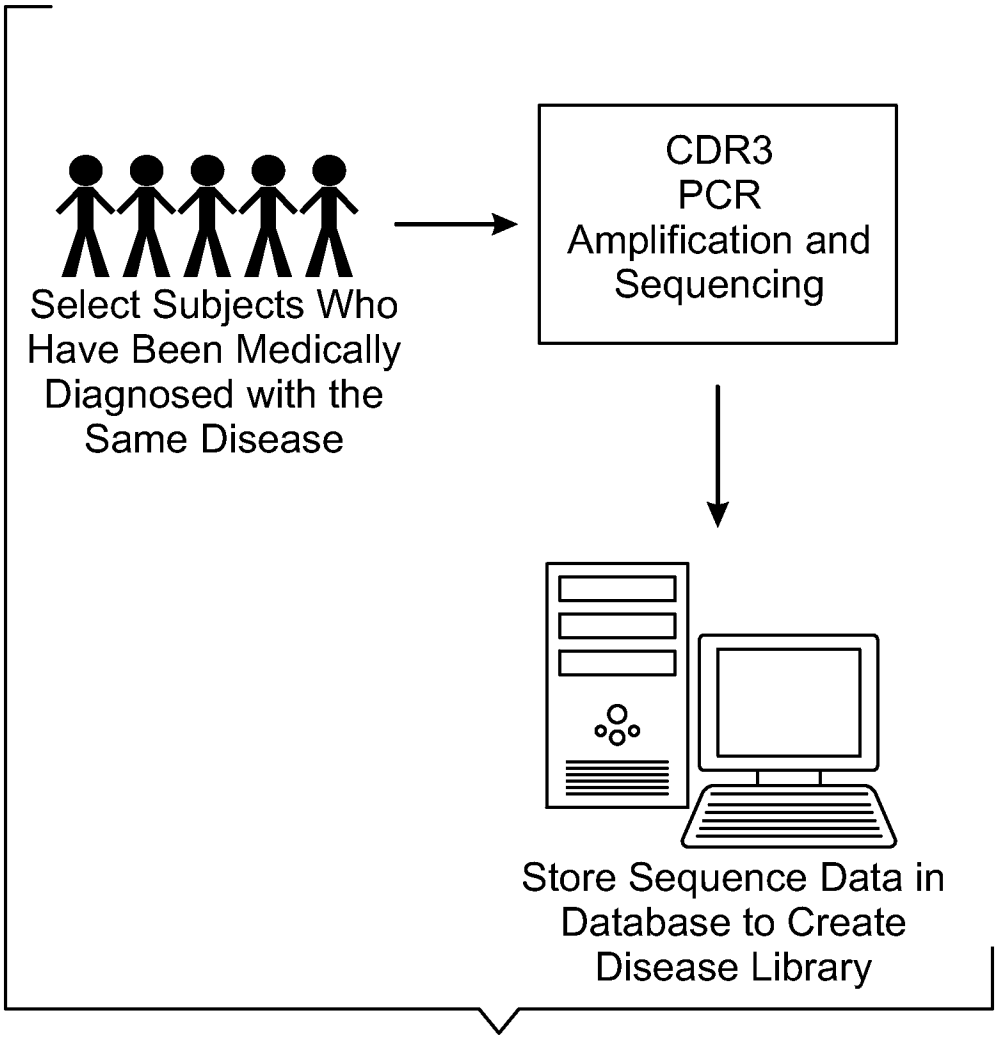

Select Subjects Who Have Been Medically Diagnosed with the Same Disease
CDR3 PCR Amplification and Sequencing
Store Sequence Data in Database to Create Disease Library

*FIG. 4*

| CDR3 | Read count | Percentage | Shared? |
|---|---|---|---|
| $CDR3_1$ | 120345 | 0.0602% | **Yes** |
| $CDR3_2$ | 1542 | 0.0008% | No |
| $CDR3_3$ | 4530 | 0.0023% | No |
| $CDR3_4$ | 8762 | 0.0044% | **Yes** |
| $CDR3_5$ | 689 | 0.0003% | No |
| $CDR3_6$ | 325 | 0.0002% | No |
| $CDR3_7$ | 8452 | 0.0042% | **Yes** |
| $CDR3_8$ | 23540 | 0.0118% | **Yes** |
| $CDR3_9$ | 3841 | 0.0019% | No |
| $CDR3_n$ | 20 | 0.0000% | No |
| Sum | | | 0.0805495% |

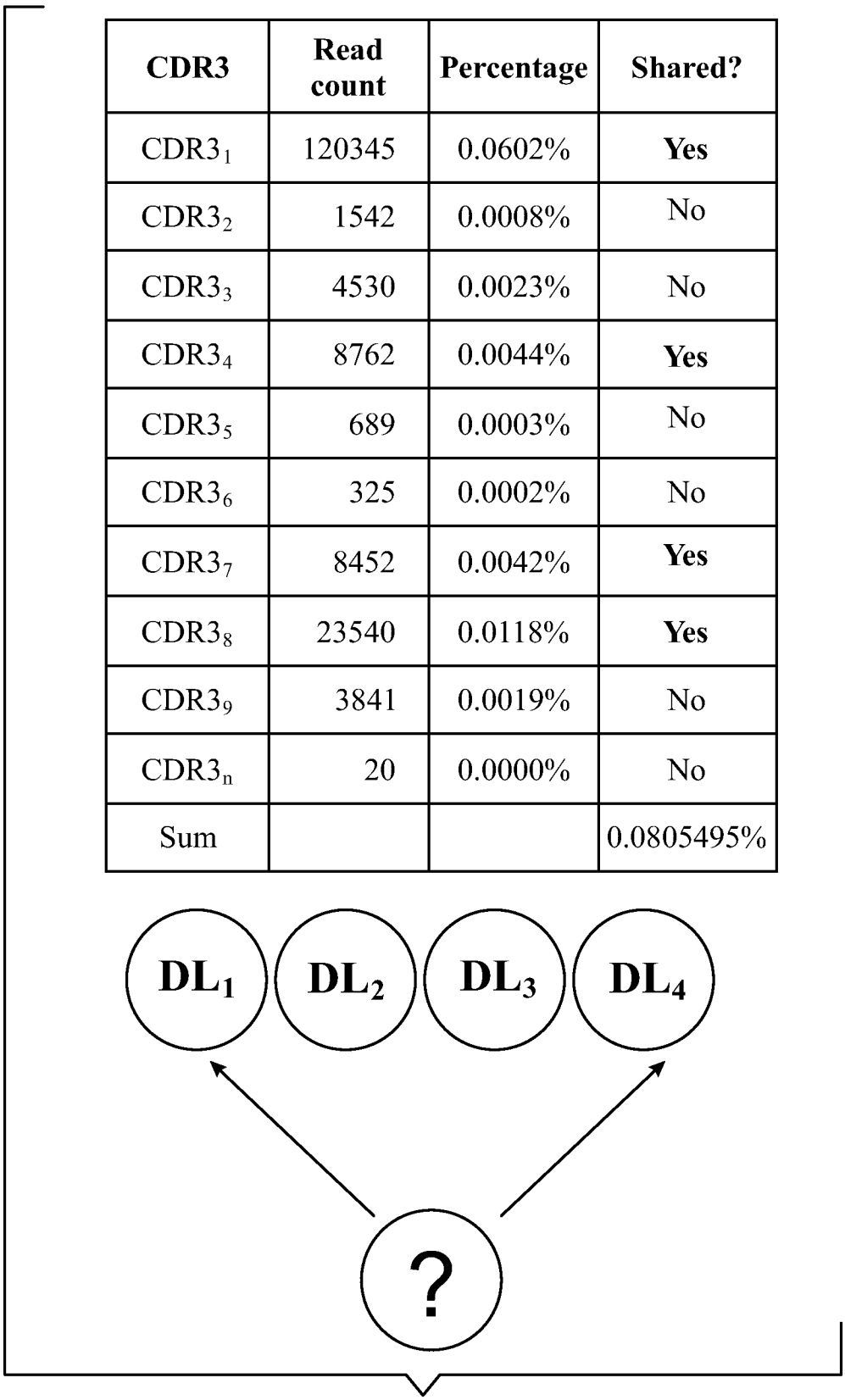
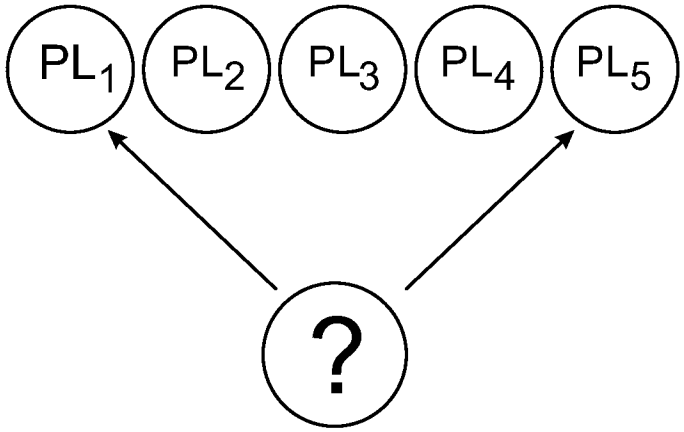



*FIG. 5*




| CDR3 | Read count | Percentage | Shared? |
|---|---|---|---|
| $CDR3_1$ | 120345 | 0.0602% | N |
| $CDR3_2$ | 1542 | 0.0008% | **Y** |
| $CDR3_3$ | 4530 | 0.0023% | **Y** |
| $CDR3_4$ | 8762 | 0.0044% | N |
| $CDR3_5$ | 689 | 0.0003% | **Y** |
| $CDR3_6$ | 325 | 0.0002% | **Y** |
| $CDR3_7$ | 8452 | 0.0042% | N |
| $CDR3_8$ | 23540 | 0.0118% | N |
| $CDR3_9$ | 3841 | 0.0019% | N |
| $CDR3_n$ | 20 | 0.0000% | **Y** |
| Sum | | | 0.003553% |

*FIG. 6*

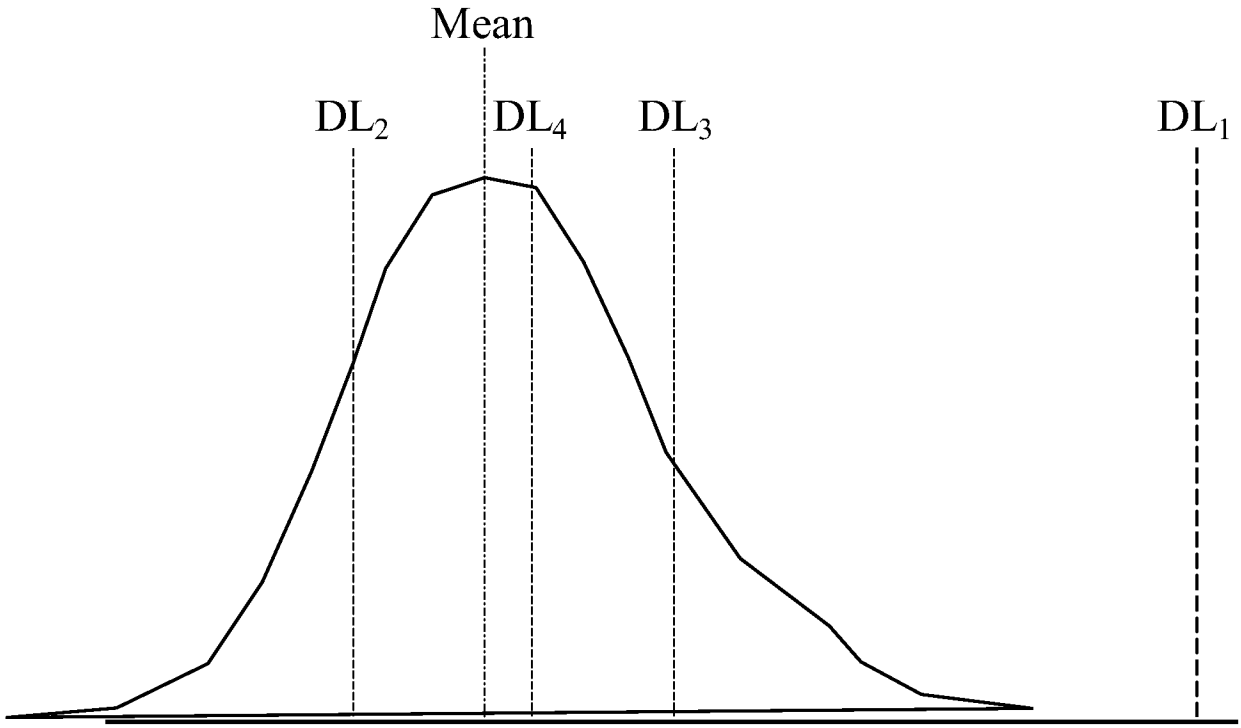



*FIG. 7*

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US14/15841

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2014.01)
USPC - 435/6.12, 6.11, 6.1, 4
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)
IPC(8): C12Q 1/68 (2014.01)
USPC: 435/6.12, 6.11, 6.1, 4

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-C,B, DE-A, DE-T, DE-U, GB-A, FR-A); Google; Google Scholar; PubMed; ProQuest; evaluating, determine, identify, changes, rearrangements, recombination, 'immune response cell population,' 'T cells,' 'B Cells,' 'RNA amplification,' 'RT-PCR,' sequencing, associating, link, connect, correlate, disease

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X<br>—<br>Y | US 2010/0021896 A1 (HAN, J) January 28, 2010; abstract; paragraphs [0007], [0009]-[0011], [0014], [0016], [0017], [0020], [0024], [0026], [0034], [0039], [0077] | 12-17<br>—<br>1-10 |
| Y | US 2012/0171725 A1 (HAN, J) July 5, 2012; paragraphs [0017], [0019] | 1-10 |
| Y | WO 2012/097374 A1 (HAN, J) July 19, 2012; paragraphs [0007], [0021], [0022] | 2, 11 |
| Y | WANG, C et al. High Throughput Sequencing Reveals A Complex Pattern Of Dynamic Interrelationships Among Human T Cell Subsets. PNAS. 26 January 2010, Vol. 107; pages 1518-1523, Supporting Information; pages 1-16; page 1518, column 1, paragraph 3, column 2, paragraphs 1-2; page 1519, column 2, paragraph 2; page 1520, column 2, paragraphs 5-6; Supporting Information, page 3, paragraph 3; Supporting Information, page 10, paragraph 1; Table 1; Table S5; figure S2. | 11 |
| A | WO 2011/140433 A2 (QUAKE, SR et al.) November 10, 2011; entire document | 11 |

☐ Further documents are listed in the continuation of Box C. ☐

* 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 |
|---|---|
| 22 April 2014 (22.04.2014) | 12 MAY 2014 |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-3201 | Authorized officer:<br>Shane Thomas<br>PCT Helpdesk: 571-272-4300<br>PCT OSP: 571-272-7774 |

Form PCT/ISA/210 (second sheet) (July 2009)

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US14/15841

**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 filed or furnished:

   a. (means)

   ☐ on paper

   ☒ in electronic form

   b. (time)

   ☒ in the international application as filed

   ☐ together with the international application in electronic form

   ☐ subsequently to this Authority for the purposes of search

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 in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments: