

METHOD FOR IDENTIFYING DISEASE-ASSOCIATED CDR3 PATTERNS IN AN IMMUNE REPERTOIRE

Cross Reference to Related Application

[0001] This application claims priority to U.S. Provisional Patent Application No. 62/130,512, entitled “Method for Identifying Disease-Associated CDR3 Patterns in an Immunorepertoire” and filed on March 9, 2015, which is incorporated herein by reference.

Field of the Invention

[0002] The invention relates to methods for recognizing disease-associated immune repertoires in human and/or animal subjects and for the use of such methods for disease diagnosis and for the study of disease processes.

Background of the Invention

[0003] The diverse antigen receptors of T and B lymphocytes are produced by somatic recombination of a limited, but large number of gene segments. These gene segments – V (variable), D (diversity), J (joining), and C (constant) – determine the binding specificity and downstream applications of immunoglobulins and T cell receptors (TCRs). The rearranged V(D)J portion of the receptor, termed the V-region, is of great interest, because it is responsible for epitope recognition. When the V(D)J is translated into an amino acid sequence, the V-region can be subdivided into several parts consisting of the leader sequence, framework (FR) 1, complementarity-determining region 1 (CDR1), FR2, CDR2, FR3, CDR3, FR4, and the C-domains.

[0004] The CDR3 is of particular interest because studies have indicated that this region is associated with antigen-specificity. Compared to normal subjects, patients with various diseases may experience quantitative and/or qualitative changes in their immune repertoire. Quantitative changes may be apparent as increases and decreases in immune repertoire diversity. Qualitative changes may present as increased sharing of disease-specific CDR3s in T or B cells.

[0005] The immune system mounts a response to various conditions, such as cancer, bacterial infections, viral infections, and fungal infections. Further, in some subjects it may actually produce a deleterious response to the tissues of the body, resulting in, for example, autoimmune disease or rejection of grafts. The types and degrees of these immune responses could, if accurately accessed, potentially be one of the most important and accurate indicators of the presence or absence of a particular disease or of undesirable immune responses.

[0006] Humans, however, are estimated to have as many as 10^{15} - 10^{25} different T cells, due to the number of possible VDJ rearrangements, n-additions, and alpha-beta chain combinations. Assuming that, for a particular disease, there are only 10^3 disease-specific CDR3s that may quantitatively (through up or down regulation) or qualitatively (through gain or loss) change, the signal-to-noise ratio in these circumstances is very weak. Therefore, conventional methods are impractical to assess this information for diagnostic purposes.

[0007] What are needed are improved methods for immune response assessment and for the development of diagnostic tests based on such immune response and the resulting immune repertoire.

Summary

[0008] In one aspect the present invention refers to a method for developing a diagnostic test for a particular disease, the method comprising the steps of amplifying and sequencing the CDR3 sequences in a sample collected from each of multiple subjects in a patient group and multiple subjects in a control group, wherein the patient group comprises subjects who have the same disease and the control group comprises subjects who are healthy, and determining the frequency of occurrence of each unique CDR3 sequence present in the samples to identify Public CDR3 sequences that are shared between at least two subjects in each of the patient group and the control group, wherein each subject in the patient group is diagnosed with the same disease and wherein each subject in the control group is healthy and not diagnosed with the disease; ranking the Public CDR3 sequences by order of frequency of occurrence; and identifying Positive Linklets from the Public CDR3 sequences, wherein a disease signature is present when a minimum number of Significant Linklets are identified in a sample, wherein Significant

Linklets are Positive Linklets that are associated to a statistically significant degree of $p < 0.05$ with subjects in the patient group relative to subjects in the control group. In another aspect the present invention refers to a method for developing a diagnostic test for a particular disease, the method comprising the steps of amplifying and sequencing the CDR3 sequences in a blood sample, collected from each of multiple subjects in a patient group and multiple subjects a control group, wherein the patient group comprises subjects who have the same disease and the control group comprises subjects who are categorized as healthy, and determining the frequency of occurrence of each unique CDR3 sequence present in the samples to identify at least about 1000 Public CDR3 sequences that are shared between at least two subjects in each of the patient group and the control group, wherein each subject in the patient group is diagnosed with the same disease and wherein each subject in the control group is healthy and not diagnosed with the disease, wherein the patient group and the control group each comprise at least 100 subjects; ranking the Public CDR3 sequences by order of frequency of occurrence; and identifying at least about 106 Positive Linklets from the Public CDR3 sequences, wherein a disease signature is present when a minimum number of Significant Linklets are identified in a sample, wherein Significant Linklets are Positive Linklets that are associated to a statistically significant degree of $p < 0.05$ with subjects in the patient group relative to subjects in the control group. The present disclosure also describes a method for developing a diagnostic test using the immune repertoire, the method comprising the steps of: (a) collecting a sample from each of multiple subjects in a patient group and a control group, wherein the patient group comprises subjects who have the same disease and the control group comprises subjects who are categorized as healthy; (b) amplifying and sequencing the immune repertoire of each subject in each of the two groups to identify each unique CDR3 sequence present in the sample and to determine the frequency of occurrence of each unique CDR3 sequence; (c) identifying CDR3 sequences that are shared between at least two subjects in each of the control group and the patient group; (d) ranking the identified CDR3 sequences by order of frequency of occurrence; (e) identifying Linklets from each group; and (f) identifying the Linklets that are associated to a statistically significant degree with the patient group to provide a disease signature.

[0009] In certain examples, the sample is peripheral blood. In other examples, the sample is tissue. In certain examples, fewer than 1,000 CDR3 sequences are identified as shared between at least two subjects. In other examples, at least 1,000 CDR3 sequences are identified as shared between at least two subjects. In certain examples, at least about 10^6 Linklets are identified from each group. In other examples, fewer than 10^6 Linklets are identified from each group.

Brief Description of the Figures

[0010] The disclosure can be better understood with reference to the following figures.

[0011] Fig. 1 is a cartoon illustrating the basic concept of the disclosed method, where polynucleotide sequence data from immune system cells is processed using software designed to sort and count the sequences, rank them by frequency numbers, generate p values, and other criteria, to produce a diagnostic signature from the Linklets designated as “Significant Linklets.”

[0012] Fig. 2 is a table illustrating the ranking of CDR3s by number of clones present in a sample. Typically, a sequencing result from one sample will generate as many as 400,000 CDR3s. Each CDR3 is associated with a read count; due to the semi-quantitative nature of the amplification method (arm-PCR), the read count also reflects the relative abundance of the clone. Software analysis removes the errors and ranks the CDR3s to produce an output file as shown in the table.

[0013] Fig. 3 is a table illustrating how Linklets are detected in the sequences obtained from blood samples. CDR3 sequences were tallied to provide a list of the CDR3 sequences that are present in the highest numbers in a blood sample. Among those CDR3 sequences, Linklets represent pairs of CDR3s that are present within the same sample – at a level that is higher than a designated cutoff level.

[0014] Fig. 4 is a representative list of some of the Linklets identified during a breast cancer study (comparison between Linklets detected in two study groups – a first group of subjects who had been diagnosed with breast cancer and a second group of subjects who were designated as healthy controls).

[0015] Fig. 5 lists Public Linklets detected in the breast cancer study, based on their p values. The top ranked Linklet, for example, is the CDR3 pair ‘ASSYSRGEEF’ and

‘ASSLGRTHQPQH’, and this Linklet was shared in 32 of the 98 breast cancer patient samples, while only 1 of the 106 control samples had it. The p value was 0. Only those Linklets with a p value < 0.05 are included in the final list that represents a breast cancer diagnostic signature. A total of 101,902 Significant Linklets were identified.

[0016] Fig. 6 is a scatter plot illustrating results for subjects in three groups: control, breast cancer, and CMV (cytomegalovirus). A receiver operating characteristic (ROC) curve analysis suggested a cutoff value of 600 DSLs. Of the 103 breast cancer patients, 98 had more than 600 Significant Linklets; of the 110 controls, only 7 had more than 600 Significant Linklets. The diagnostic sensitivity and specificity were therefore 95% and 93%, respectively. When 188 non-breast cancer samples (from patients enrolled in a CMV study) were studied against the breast cancer disease signature, only 3 samples were false positive with more than 600 Significant Linklets, giving a specificity of 98.4%.

Detailed Description

[0017] The present disclosure generally pertains to a method for developing diagnostic tests that are based on the immune response and the resulting immune repertoire. The presently disclosed method increases the signal and reduces the background to allow the identification of shared CDR3s that can be used to produce a disease signature. The presently disclosed method may be used to develop a diagnostic test for different diseases including, but not limited to, cancer, autoimmune disease, inflammatory disease and infectious disease.

[0018] As used herein, “disease” includes diagnosed disease and other disruptions, diagnosed and undiagnosed, to the normal health of a subject.

[0019] As used herein, “healthy” means not currently exhibiting symptoms of, and not currently diagnosed with, a disease.

[0020] As used herein, an “immune repertoire” comprises the functionally diverse T and B cells of a subject.

[0021] As used herein, a “Linklet” is a pair of unique CDR3s that are present in the same sample. When two or more people share a particular Linklet, it is a “Public Linklet.” If a Linklet is only detected in one subject, it is a “Private Linklet.” Public Linklets come

from Public CDR3s (i.e., CDR3s that are detected in more than one subject). Generally speaking, each subject's repertoire is largely "private" and only a small percent of that subject's immune repertoire represents shared CDR3s. Public Linklets are therefore present at a much lower level than are Private Linklets, a fact that makes identification of disease signatures more difficult. It is therefore important to utilize an approach that reduces the background to allow identification of the significant CDR3 repertoire that constitutes one or more disease signatures.

[0022] As used herein, "sample" comprises blood and tissue. In certain embodiments, blood is peripheral blood collected from a subject. In certain embodiments, tissue is a biopsy obtained from a subject.

[0023] As used herein, "subject" means a human or animal.

[0024] In certain embodiments, the presently disclosed method reduces the background through the use of "Positive Linklets." When two CDR3 sequences, A and B, are sequenced, quantitative information is also obtained, represented by the read counts. If the immune repertoire amplification method used is semi-quantitative (arm-PCR), and if CDR3-A is expressed in a sample at a higher level than CDR3-B, more sequence read counts will be obtained for A than for B. In such a scenario, the A-B pair is designated as a Positive Linklet, whereas the B-A pair would be designated as a Negative Linklet. In certain embodiments of the presently disclosed method, only the Positive Linklets are used for further analysis. Use of Positive Linklets enriches the diagnostic signal, because it helps to filter out experimental noise.

[0025] Biologically, more than one antigen or epitope is generally associated with a particular disease. Therefore, relevant T and B cell receptors would generally appear in patients' samples as clusters. The quantitative information provided by Positive Linklets (and Negative Linklets) may reflect the disease-specific antigen expression profile.

[0026] Experimental procedures may introduce additional "noise" in the data generated. For example, it is a common practice to pool many samples from different subjects into one sequencing run to reduce the cost (the immune repertoire being amplified separately using barcoded primers). However, if a CDR3 is very dominant in one sample, it will appear on the sequencing chip multiple times. That dominant clone will have a 1/8,000 chance to be assigned to the wrong barcode and be "shared" by all the other samples in

the same run. If CDR3 is used as the basic analytical unit, these “contaminated” sequences would be considered as biologically shared and be used as diagnostic signals. Using Linklets allow these noises to be filtered out, because those incorrectly assigned CDR3s are usually at very low frequency, and the likelihood that they will be part of one or more Positive Linklets is reduced (with a higher likelihood that they will form Negative Linklets). By considering only Positive Linklets, the noise can be filtered out. Also, if only the top ranked CDR3s from a sample are used (such as, for example 5,000, or between 1,000 and 50,000 of the top ranked CDR3s), those incorrectly assigned CDR3s usually will not be considered, due to their low frequencies.

[0027] When a group of Public Linklets are found to be associated to a higher degree with a group of subjects who have a particular disease in common, those Public Linklets can be treated as disease-specific Linklets, or “Significant Linklets.” A group of Significant Linklets associated with a particular disease can therefore constitute a “disease signature.” Therefore, if a subject’s sample is found to have statistically significant overlap with the disease signature, a diagnosis of such disease can be made for that subject.

[0028] The presently disclosed method comprises the following steps: (1) gathering samples from subjects assigned to a patient group and a control group; (2) amplifying and sequencing an immune repertoire for each sample; (3) identifying the unique CDR3 sequences from each sample’s immune repertoire; (4) tallying the number of times an individual (unique) CDR3 sequence is detected in the immune repertoire, thereby identifying those clones that are dominant (determined by ranking them in order of highest frequency of occurrence to lowest frequency of occurrence); (5) comparing the immune repertoires of the subjects to identify CDR3s that are shared between at least two subjects (“Public CDR3s”); (6) ranking the Public CDR3s based on their frequencies of occurrence; (7) generating a list of Positive Linklets from the top-ranked CDR3s; (8) filtering out the Private Linklets and retaining the Public Linklets; and (9) identifying Public Linklets that are associated with patients in the target disease group, but not with the control group.

[0029] In certain embodiments, at least about 100 subjects are assigned to each group. In certain embodiments, the immune repertoire is amplified using the arm-PCR method

(described in WO2009/124293). In certain embodiments, the top 5,000 clones are identified as dominant. In certain embodiments, the list of Positive Linklets includes the top-ranked 1,000 to 20,000 CDR3s. In certain embodiments, the confidence value for Public Linklets associated with patient in the target disease group, but not with the control group, is $p < 0.05$.

[0030] The list of Public Linklets that are associated with patients (disease associated Linklets, or DSLs), but not with the control group, constitutes a group that is designated as “Signature Linklets.” In certain embodiments, a signature may be obtained by analyzing about 100 patients and an equal number of controls. A cutoff Disease Signature Linklet (DSL) value is then determined. Unknown samples may be tested by sequencing followed by counting the DSLs. If the DSL number meets or exceeds the cutoff, a diagnosis is made for a particular disease.

[0031] For example, in certain embodiments to obtain the sequence data for analysis, whole blood from a subject (e.g., human or animal, disease group or control (healthy)) is treated with Ficoll® to extract peripheral blood mononuclear cells, or PBMCs, to get the highest concentration of lymphocytes. Each type of lymphocyte has a specific identifier called a CD marker, or cluster of differentiation marker, which is numbered. For example, cytotoxic T-cells have a CD8 marker, and helper T-cells have a CD4 marker. Magnetic beads, which have been labeled with a specific anti-CD marker, can be added to the cell suspension. After applying the column to a magnetic field, the bead bound cells will be trapped, or positively-selected, while allowing the other cell types to flow through. The flow through, or negatively-selected cell suspension, can be used to further isolate other cell populations in downstream applications. In other embodiments, the sample may be tissue, which can be processed, using methods known in the art, to isolate lymphocytes.

[0032] Since there are sub-populations within certain populations of T cells (e.g., regulatory T cells are a subpopulation of helper T-cells), if sub-populations need to be separated, release reagents can be added to the CD4-bead-bound cells to release the bead, so that another magnetic bead can bind to the cell. In the case of regulatory T-cells, for example, a CD25⁺ selection microbead can be added to the cell suspension to extract the regulatory T-cell population from the helper T-cell population.

- [0033] Polynucleotide isolation (RNA or DNA) can be performed by means known to those of skill in the art (see, e.g., Murray, BMC Res Notes. 2013 Nov 1;6:440). Amplification of sequences may be performed using the method described in WO2009/124293 (arm-PCR), which provides the sensitivity and specificity that is necessary to achieve superior results in the presently disclosed method.
- [0034] Sequencing may also be performed using methods known in the art. Given the numbers of sequences that must be determined, high-throughput sequencing methods are generally employed, such as, for example, Illumina's Next-Generation Sequencing, using Illumina sequencing primers.
- [0035] Large amounts of data are generated and must be analyzed and manipulated as a result of the sequencing, tallying the number of times a particular sequence (representing an individual clone) occurs, ranking the clones in order based on frequency of occurrence, and other analyses described herein. This is most conveniently and effectively performed using sequence data analysis programs. One such program is CDR3 Algebra, which does the sorting, ranking, and pairing for the researcher. Statistical analysis, such as calculation of p values, can also be performed using such programs.
- [0036] The presently disclosed method lends itself to the development of diagnostic tests for a variety of diseases including, but not limited to, cancer, autoimmune disease, bacterial infections, viral infections, and fungal infections, thereby giving researchers and clinicians a valuable tool for the diagnosis and study of a disease of interest.
- [0037] The presently disclosed method can be further described by means of the following non-limiting examples.

Examples

Isolation of Peripheral Blood Mononuclear Cells (PBMCs) from Whole Blood

- [0038] Whole blood from healthy subjects (control group) and patients previously diagnosed with breast cancer (patient group) was diluted with PBS buffer at 2-4x the original volume. 10 mL of whole blood collected in sodium heparin was transferred to a 50 mL conical tube and diluted with buffer to the 35 mL line. Diluted cell suspension (35 mL) was carefully layered over 15 mL of Ficoll-Paque® in a separate 50 mL conical

tube. The tube was centrifuged at 400 x g for 30 minutes at 20 degrees Celsius in a swing bucket rotor with no brake.

[0039] The upper layer containing PBS buffer and plasma was carefully aspirated to remove it. The cloudy mononuclear cell layer was carefully transferred to a fresh 50 mL conical tube. The tube was then filled with buffer to the 50 mL mark and centrifuged at 300 x g for 20 minutes at 20 degrees Celsius. The clear supernatant was removed and the cell pellet was re-suspended in 8 mL of buffer.

Isolation of Monocytes from Isolated PBMCs

[0040] Cells were counted using a hemocytometer and the sample centrifuged at 300 x g for 10 minutes at room temperature. The supernatant removed by aspiration. Cells were resuspended in 80 μ L of buffer per 10^7 cells.

[0041] Twenty microliters of CD14 Microbeads were added per 1×10^7 cells, and mixing was performed by gently pipetting up and down. The microbead/cell mixture was incubated at 4°C for 15 minutes. Cells were washed by adding 2 mL of buffer per 1×10^7 cells and were then centrifuged at 300 x g for 10 minutes.

[0042] The supernatant was aspirated completely and was resuspended in buffer (10^8 cells in 500 μ L of buffer). An LS magnetic column was placed on the magnet and washed with 3 mL of buffer. Flow-through buffer was discarded. Cell suspension was applied to the column and unlabeled cells that pass through were collected in a labeled 15 mL conical tube. The column was washed 3 times with 3 mL of buffer, with new buffer added only when the column reservoir was empty.

[0043] A new, clean 15 mL conical tube labeled “Monocyte” was placed under the column and the column was removed from the magnet. Buffer (5 mL) was pipetted into the column and the magnetically labeled cells were immediately flushed out by firmly pushing the plunger into the column.

[0044] Both tubes were centrifuged for 10 minutes at 300 x g, and the supernatant completely aspirated. For the tube labeled “Monocyte”, the cells were re-suspended in 2 mL of buffer. Twenty microliters were pipetted out to be used for the cell counting protocol, and the tube centrifuged at 300 x g for 10 minutes. Cells were resuspended in 500 μ L of RNeasy Protect® and stored at 4°C for later extraction of RNA. For the tube labeled “CD14-”, the cells were re-suspended in 80 μ L of buffer per 10^7 cells.

RNA Extraction

[0045] Cells were centrifuged for 3 minutes at 3,000 rpm at 20°C, supernatant was removed, and the cell pellet was loosened by flicking the tube. BME Buffer (350 µL) was added to the sample, and the cell pellet was dissolved completely by vortexing.

[0046] The sample was transferred to a QIAshredder column and homogenized by centrifuging for 2 minutes at 10,000 rpm. The column was discarded, and the flow through was saved. Ethanol (70%, 350 µL) was added to the flow through and the sample was mixed by pipetting. The sample (700 µL) was transferred to an RNeasy® spin column and placed in a 2 ml collection tube. The sample was centrifuged for 15 seconds at 10,000 rpm. Flow through was discarded. In cases where there was more than 700 µL of sample, this step was repeated using the same column.

[0047] 700 µL of Buffer RW1 was added to the spin column and the sample was centrifuged at 10,000 rpm for 15 seconds, discarding the flow through. 500 µL of Buffer RPE was added to the spin column, and the sample was centrifuged at 10,000 rpm for 15 seconds, discarding the flow through.

[0048] 500 µL of Buffer RPE was added to the spin column and the sample was centrifuged for 2 minutes at 10,000 rpm. The spin column was placed in a new 2 mL collection tube and was centrifuged for 1 minute at 10,000 rpm to dry the column membrane. The spin column was placed in a new 1.5 mL collection tube. 25 µL of RNase-free water was added to all samples except for samples containing isolated regulatory T cells. To regulatory T cell samples, 20 µL of RNase-free water was added. The sample was allowed to sit at room temperature for 1 minute. The sample was centrifuged for 1 minute at 10,000 rpm and the column was discarded.

Amplification of CDR3 Sequences Using Polymerase Chain Reaction

[0049] PCR amplification of CDR3 sequences was performed using the arm-PCR method disclosed in WO2009/124293 (Han). A minimum of 100 ng of RNA or gDNA (depending on the reagent system selected) with a 260/280 of 1.8 or greater is generally recommended as the starting material to obtain the best diversity of the arm-PCR immune repertoire library. During the first round of PCR, nested gene specific primers targeting each of the V and J (or C) genes were used. The forward primers, F_o (forward-out) and F_i (forward-in), targeted the V genes. The reverse primers, R_o (reverse-out) and R_i (reverse-

in), targeted each of the J or C genes. The F_i and R_i primers also included sequencing adaptors B and A, respectively, for the Illumina® platforms (HiSeq, MiSeq and GAIIx) for paired-end sequencing. The second round of PCR was carried out using communal (common) primers B and A. After gel purification, the resulting product was ready for high throughput sequencing with the Illumina® platforms. The first round of PCR introduced barcodes and sequencing primers into the PCR products.

[0050] The exponential phase of the amplification was achieved by the communal primers in the second round of PCR; therefore, the target immune repertoire was amplified evenly and semi-quantitatively, without introducing additional amplification bias.

Identification of Breast Cancer Signature

[0051] A total of 213 samples were collected, including 103 from breast cancer patients and 110 from controls. A total of 14,666,172 CDR3s were identified from the 213 samples, averaging 68,855 CDR3s from each sample. 8,301,648 unique CDR3s were found from the 213 samples. After removing the private CDR3s, a total of 98,076 public (i.e., shared by at least two subjects) and dominant (i.e., ranked within the top 5,000 CDR3s in each sample) CDR3s were identified from the 213 samples, using iRepertoire (Huntsville, Alabama USA) software available through the company website (e.g., CDR3 Algebra). A total of 287,198,206 Positive Linklets were generated from the 213 samples, averaging 1,003,236 Linklets from each sample. After removing Private Linklets, 16,921,605 Linklets remained that were shared with at least one other person. For each shared Linklet, a p value was obtained to identify those preferentially shared among patients. A total of 117,069 Linklets were identified as Significant Linklets with $p < 0.05$, providing a “signature” for the diseases. A total of 6,171 CDR3s contributed to the 117,069 disease signature Linklets. Using a cutoff value of 600 Significant Linklets, 95% of breast cancer could be diagnosed, with 93.6% of specificity. When 188 non-breast cancer samples were studied, only three samples were false positive (having more than 600 DSLs), giving a specificity of 98.4%.

[0052] The presently disclosed method increases the signal and reduces the background to allow the identification of shared CDR3s that can be used to produce a disease signature, which otherwise is not possible using conventional methods. As a result, the

presently disclosed method has the benefit of allowing the development of a diagnostic test for different diseases including, but not limited to, cancer, autoimmune disease, inflammatory disease and infectious disease.

[0053] This application references various publications. The disclosures of these publications, in their entireties, are hereby incorporated by reference into this application to describe more fully the state of the art to which this application pertains. The references disclosed are also individually and specifically incorporated herein by reference for material contained within them that is discussed in the sentence in which the reference is relied on.

[0054] The methodologies and the various embodiments thereof described herein are exemplary. Various other embodiments of the methodologies described herein are possible.

CLAIMS

Now, therefore, the following is claimed:

1. A method for developing a diagnostic test for a particular disease, the method comprising the steps of:

amplifying and sequencing the CDR3 sequences in a sample collected from each of multiple subjects in a patient group and multiple subjects in a control group, wherein the patient group comprises subjects who have the same disease and the control group comprises subjects who are healthy, and

determining the frequency of occurrence of each unique CDR3 sequence present in the samples to identify Public CDR3 sequences that are shared between at least two subjects in each of the patient group and the control group, wherein each subject in the patient group is diagnosed with the same disease and wherein each subject in the control group is healthy and not diagnosed with the disease;

ranking the Public CDR3 sequences by order of frequency of occurrence; and

identifying Positive Linklets from the Public CDR3 sequences,

wherein a disease signature is present when a minimum number of Significant Linklets are identified in a sample,

wherein Significant Linklets are Positive Linklets that are associated to a statistically significant degree of $p < 0.05$ with subjects in the patient group relative to subjects in the control group.

2. The method of claim 1, wherein the sample is blood.

3. The method of claim 1, wherein the sample is tissue.

4. The method of claim 1, wherein at least about 1,000 CDR3 sequences are identified that are shared between at least two subjects in each of the control group and the patient group.

5. The method of claim 1, wherein at least about 10^6 Positive Linklets from each of the patient group and control group are identified.

6. The method of claim 1, wherein the disease is breast cancer.

7. The method of claim 6, wherein the minimum number of Significant Linklets is at least about 500.

8. A method for developing a diagnostic test for a particular disease, the method comprising the steps of:

amplifying and sequencing the CDR3 sequences in a blood sample, collected from each of multiple subjects in a patient group and multiple subjects a control group, wherein the patient group comprises subjects who have the same disease and the control group comprises subjects who are categorized as healthy, and

determining the frequency of occurrence of each unique CDR3 sequence present in the samples to identify at least about 1000 Public CDR3 sequences that are shared between at least two subjects in each of the patient group and the control group, wherein each subject in the patient group is diagnosed with the same disease and wherein each subject in the control group is healthy and not diagnosed with the disease, wherein the patient group and the control group each comprise at least 100 subjects;

ranking the Public CDR3 sequences by order of frequency of occurrence; and

identifying at least about 10^6 Positive Linklets from the Public CDR3 sequences,

wherein a disease signature is present when a minimum number of Significant Linklets are identified in a sample,

wherein Significant Linklets are Positive Linklets that are associated to a statistically significant degree of $p < 0.05$ with subjects in the patient group relative to subjects in the control group.