

DETECTION AND MEASUREMENT OF TISSUE-INFILTRATING LYMPHOCYTES**CROSS REFERENCE**

[0001] This application claims priority to co-pending U.S. provisional application number 61/570,192 filed 13 December 2011, which application is incorporated herein in its entirety by reference.

BACKGROUND OF THE INVENTION

[0002] The numbers and ratios of different lymphocyte subsets infiltrated into a disease-affected tissue, such as a solid tumor, often bears on the prognosis of the disease, e.g. Deschoolmeester et al, BMC Immunology, 11:19 (2010); Ohtani, Cancer Immunity, 7: 4 (2007); Yu et al, Laboratory Investigation, 86: 231-245 (2006); Diederichsen et al, Cancer Immunol. Immunother., 52: 423-428 (2003); and the like. Unfortunately, measurement of such quantities using available technologies, such as immunohistochemistry or flow cytometry, is difficult, labor intensive, and not amenable for routine deployment.

[0003] Separately, there has been more and more interest in the use of large-scale DNA sequencing in diagnostic and prognostic applications as the per-base cost of DNA sequencing has dropped. For example, profiles of nucleic acids encoding immune molecules, such as T cell or B cell receptors, or their components, contain a wealth of information on the state of health or disease of an organism, so that the use of such profiles as diagnostic or prognostic indicators has been proposed for a wide variety of conditions, e.g. Faham and Willis, U.S. patent publication 2010/0151471; Freeman et al, Genome Research, 19: 1817-1824 (2009); Boyd et al, Sci. Transl. Med., 1(12): 12ra23 (2009); He et al, Oncotarget (March 8, 2011).

[0004] It would be highly useful to the medical and scientific fields if the improvements in high throughput nucleic acid sequencing could be put to use to provide a more convenient and more effective assay for measuring tissue-infiltrating lymphocytes (TILs).

SUMMARY OF THE INVENTION

[0005] The present invention is drawn to methods for measuring numbers, levels, and/or ratios of cells, such as lymphocytes, infiltrated into a solid tissue, such as a tumor, and to making patient prognoses based on such measurements. The invention is exemplified in a number of

implementations and applications, some of which are summarized below and throughout the specification.

[0006] In one aspect, the invention is directed to methods for identifying lymphocytes that have infiltrated a solid tissue comprising the following steps: (a) sorting into one or more subsets a sample of lymphocytes from an accessible tissue of an individual; (b) generating clonotype profiles for each of the one or more subsets of lymphocytes from the accessible tissue; (c) generating at least one clonotype profile from at least one sample of the solid tissue; and (d) detecting lymphocytes of each subset in the solid tissue from their respective clonotypes.

[0007] In another aspect, the invention is directed to methods for determining a prognosis from a state of lymphocyte infiltration into a solid tumor of a patient, wherein such method comprises the steps of: (a) sorting into one or more subsets a sample of lymphocytes from peripheral blood of the patient; (b) generating clonotype profiles for each of the one or more subsets of lymphocytes from the peripheral blood; (c) generating at least one clonotype profile from at least one sample of the solid tumor; and (d) determining numbers, levels, and/or ratios of lymphocytes of each of the one or more subsets. In one embodiment, the state of lymphocyte infiltration into a solid tumor means the number, levels, and/or ratios of lymphocytes of selected functional subset within a solid tumor. In some embodiment, the state of lymphocyte infiltration into a solid tumor may also include a spatial distribution of such values within or adjacent to a solid tumor.

[0008] These above-characterized aspects, as well as other aspects, of the present invention are exemplified in a number of illustrated implementations and applications, some of which are shown in the figures and characterized in the claims section that follows. However, the above summary is not intended to describe each illustrated embodiment or every implementation of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention is obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0010] FIG. 1 illustrates diagrammatically steps of one embodiment of the invention.

[0011] FIGS. 2A-2C show a two-staged PCR scheme for amplifying TCR β genes.

[0012] FIG. 3A illustrates details of determining a nucleotide sequence of the PCR product of Fig. 2C. FIG. 3B illustrates details of another embodiment of determining a nucleotide sequence of the PCR product of Fig. 2C.

[0013] FIG. 4A illustrates a PCR scheme for generating three sequencing templates from an IgH chain in a single reaction. FIGS. 4B-4C illustrates a PCR scheme for generating three sequencing templates from an IgH chain in three separate reactions after which the resulting amplicons are combined for a secondary PCR to add P5 and P7 primer binding sites. Fig. 4D illustrates the locations of sequence reads generated for an IgH chain. Fig. 4E illustrates the use of the codon structure of V and J regions to improve base calls in the NDN region.

DETAILED DESCRIPTION OF THE INVENTION

[0014] The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of molecular biology (including recombinant techniques), bioinformatics, cell biology, and biochemistry, which are within the skill of the art. Such conventional techniques include, but are not limited to, sampling and analysis of blood cells, nucleic acid sequencing and analysis, and the like. Specific illustrations of suitable techniques can be had by reference to the example herein below. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV); *PCR Primer: A Laboratory Manual*; and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press); and the like.

[0015] In one aspect, the invention is directed to methods of determining the types and numbers of lymphocytes infiltrated into a solid tissue, such as a tumor, a tissue affected by an autoimmune disease, a tissue affected by graft versus host disease (GVHD), a normal tissue, or the like. Although solid tissues of interest are usually disease-affected solid tissue, in some embodiments, the levels and/or numbers and/or ratios of different subsets of lymphocytes in normal tissues may also be used to determine states of health and/or propensities of an individual to contract a disease or condition.

[0016] An outline of one embodiment of the invention is shown in Fig. 1. Clonotypes of an individual's lymphocytes are determined from a readily accessible tissue (100), such as peripheral blood. Optionally, minimal sample preparation steps (102) may be implemented, such as isolating peripheral blood mononuclear cells (PBMCs). From such a sample, lymphocytes are sorted (104) into subsets, $L_1, L_2, \dots, L_K$ (106), which usually correspond to lymphocytes with

distinct biological functions; such subsets are sometimes referred to herein as “functional subsets” of lymphocytes. Usually, sorting is based on the presence or absence of one or more molecular markers characteristic of such functionally distinct subsets. Such markers may be cell surface markers or intracellular markers. In one embodiment, such markers are cell surface markers. Exemplary cell surface markers include, but are not limited to, CD3, CD4, CD8, CD19, CD20, CD25, CD45RO, CD117, CD127, and the like.

[0017] Lymphocyte subsets of interest include, but are not limited to, B cells; T cells; cytotoxic T cells; helper T cells; regulatory T cells; Th1 T helper T cells; Th2 helper T cells; Th9 helper T cells; Th17 helper T cells; Tfh helper T cells; antigen-specific T cells; and antigen-specific B cells. Whenever a solid tissue is a solid tumor, of particular interest are the subsets of cytotoxic T cells and regulatory T cells.

[0018] Some subsets may include members of other subsets, either because of overlap, e.g. due to an inefficient sorting technique, or because members of a second subset may be wholly contained in (or nested in) a larger first subset; for example, the subset of T cells includes cytotoxic T cells and helper T cells as two wholly contained subsets. Likewise, the subset of helper T cells includes several other wholly contained subsets, as noted. Typically, cells of the nested subsets (i.e. subsets of subsets) are obtained by using additional markers characteristic of such subsets. Subsets of lymphocyte are usually identified functionally and/or by molecular markers using conventional assays, often with commercially available markers and kits (e.g. BD Biosciences, San Jose, CA). Markers characteristic for several lymphocytes of interest are as follows: CD4 for helper T cells; CD8 for cytotoxic T cells; CD4, CD25 and low expression of CD127 for regulatory T cells (or alternatively, CD4, CD25 and intracellular expression of FoxP3 for regulatory T cells); and CD45RO⁺, CCR7⁻, CD28⁻, CD27⁻, CD8⁺ for memory effector T cells; and the like (where the “+” and “-” symbols are used as conventional in immunology literature; that is, to indicate high expression and low (or absent) expression, respectively). Antibody probes are commercially available for isolating such subsets by sorting techniques, e.g. FACS, described below. As mentioned above, in some embodiments, the presence, absence and/or levels of lymphocytes of such subsets provide prognostic information, such as the duration of survival of a patient undergoing cancer therapy. The following Table summarizes surface and intracellular markers for identifying lymphocytes subsets in accordance with the invention. Different embodiments of the invention include identifying clonotypes of different combinations of lymphocyte subsets of the Table.

Table I
Exemplary Molecular Markers for Lymphocyte Subsets
Useful for FACS Isolation*

Subset	Cell Surface Markers	Intracellular Markers
B cells	Fc receptors, CD19+, CD20+, CD21+, CD22+, CD23+	
T cells	CD3+, CD4+, CD8+	
cytotoxic T cells	CD3+, CD4-, CD8+	
helper T cells	CD3+, CD4+, CD8-	
Th1 helper T cells	CD4+, CXCR3	IFN- γ , IL-2, IL12, IL18, IL-27
Th2 helper T cells	CD4+, CCR4, Crth2	IL-4, IL-, IL-33
Th9 helper T cells	CD4+	IL-4, TGF- β
Th17 helper T cells	CD4+, CCR6	IL-17A, IL17F, IL-21, IL-22, IL-26, TNF, CCL20
Tfh helper T cells	CD4+, CXCR5	IL-12, IL-6
regulatory T cells	CD4+, CD25+, CD127 ^{-/low}	FoxP3
antigen-specific B cells	BCRs via tetramer technology	
antigen-specific T cells	TCRs via tetramer technology	

* Not meant to be an exclusive or exhaustive list.

[0019] Cell sorting based on surface markers may be carried out by one or more technologies including, but not limited to, fluorescence-activated cell sorting (FACS), magnetically-activated cell sorting (MACS), panning, resetting, and the like, which typically employ antibodies or other reagents that specifically recognize and bind to the cell surface features of interest. In one aspect, cell sorting based on intracellular markers also may be carried out using FACS by fixing and permeabilizing cells, followed by staining, e.g. with a labeled antibody specific for the intracellular marker, for example, as disclosed in Pan et al, PlosOne, 6(3): e17536 (2011). Such sorting technologies and their applications are disclosed in the following exemplary references: Recktenwald et al, editors, “Cell Separation Methods and Applications” (Marcel Dekker, 1998); Kearse, editor, “T Cell Protocols,” Methods in Molecular Biology, Vol. 134 (Springer, 2000); Miltenyi et al, Cytometry, 11: 231-238 (1990); Davies, chapter 11, “Cell sorting by flow cytometry,” in Macey, Editor, Flow Cytometry: Principles and Applications (Humana Press, Totowa, NJ); and the like. Of particular interest for the invention is sorting lymphocytes into

subsets of interest using FACS, e.g. using a commercially available instrument and manufacturer's protocols and kits, such as a BD Biosciences FACS Aria III or a BD Biosciences Influx (BD Biosciences, San Jose, CA). Using FACS to isolate regulatory T cell subsets is specifically disclosed in Boyce et al, "Human regulatory T-cell isolation and measurement of function," BD Bioscience Application Note (March, 2010), which is incorporated by reference. Sorting or isolating lymphocytes based on antigen-specificity of either T cell receptors or B cell receptors may be carried out using FACS, or FACS in combination with other technologies, such as MACS. Guidance for using such technologies for sorting and/or isolating antigen-specific T cells or B cells is disclosed in the following exemplary references, which are incorporated by reference: Thiel et al, Clin. Immunol., 111(2): 155-161 (2004); Newman et al, J. Immunol. Meth., 272: 177-187 (2003); Hoven et al, J. Immunol. Meth., 117(2): 275-284 (1989); U.S. patents 5,213,960 and 5,326,696; Moody et al, Cytometry A, 73A: 1086-1092 (2008); Gratama et al, Cytometry A, 58A: 79-86 (2004); Davis et al, Nature Reviews Immunology, 11: 551-558 (2011); U.S. patents 8,053,235 and 8,309,312; Lee et al, Nature Medicine, 5(6): 677-685 (1999); Altman et al, Science, 274: 94-96 (1996); Leisner et al, PLoSOne 3(2): e1678 (2008); "Pro5 MHC Pentamer Handbook," (ProImmune, Ltd., United Kingdom, 2012); and like references.

[0020] In one embodiment, successive samples of cells from an accessible tissue (such as peripheral blood) may be sorted into two populations: (i) a single defined subset (such as CD8⁺ lymphocytes; CD4⁺, CD25^{+(high)}; and CD127^(low); or the like) and (ii) all other cells. Population (i) is collected and analyzed, e.g. by extracting nucleic acids, amplifying recombined DNA or RNA sequences, sequencing them, and generating a clonotype profile. This procedure may be repeated for as many subsets as desired, using different subset-specific probes.

[0021] Returning to Fig. 1, DNA or RNA is extracted from each of the sorted subsets and clonotype profiles (108) are generated for each subset using the techniques described more fully below. The clonotype profiles provide a list of clonotype sequences in each subset. In one embodiment, the number of lymphocytes sorted is sufficiently large that substantially every T cell with a distinct clonotype may be identified in the clonotype profiles. As discussed more fully below, in some embodiments, since the identification includes a sampling, "substantially every cell with a distinct clonotype" means every clonotype at a given frequency or above, e.g. .0001, is determined with a probability of ninety percent, or ninety-five percent, or the like. In accordance with the invention, the clonotype information of the lymphocyte subsets is used to identify the presence, absence, numbers, and/or levels of cells of the various subsets of lymphocytes that have infiltrated into a less accessible tissue, such as a solid tumor specimen or biopsy, or tissue involved in an autoimmune disease (110). This is accomplished by extracting

DNA or RNA from specimen (110) and generating (112) a clonotype profile (114). Each clonotype of profile (114) can then be associated with a lymphocyte subset (116) by looking up the sequence of such clonotype in the subset-specific clonotype profiles (108); thus, by making such an association the subsets whose members that have infiltrated the solid tissue are identified. Moreover, because of the large diversity of clonotypes, counting the clonotypes of each subset gives a good approximation of the number of lymphocytes from the subset that have infiltrated into the specimen. If the volume of the specimen is known or determinable, then the density of lymphocytes from the subset may be obtained.

[0022] In some embodiments, such as where an inaccessible tissue is a solid tumor, one or more samples of the tumor may be taken, either before or after excision, or surgical removal. Samples taken prior to tumor removal may be obtained using needle aspirations, or other conventional techniques. In some embodiments, multiple samples are obtained, for example, to determine a spatial distribution of lymphocyte subsets within an inaccessible tissue. In some embodiments, at least two samples may be taken, at least one from the surface or exterior portion of an inaccessible tissue, and at least one from the interior of the inaccessible tissue. As noted above, in some embodiments, the inaccessible tissue is a solid tumor that has been removed from a patient, such as illustrated by specimen (110) in Fig. 1. Samples from specimen (110) may be obtained after excision and after fixation. Generating clonotype profiles from fixed tissue samples is described more fully below.

[0023] In accordance with the invention, the embodiment of Fig. 1 may be implemented with the following steps: (a) sorting into one or more subsets a sample of lymphocytes from an accessible tissue of an individual; (b) generating clonotype profiles for each of the one or more subsets of lymphocytes from the accessible tissue; (c) generating a clonotype profile from a sample of the solid tissue; and (d) detecting lymphocytes of each subset in the solid tissue from their respective clonotypes. In some embodiments, step (a) may be implemented by separating lymphocytes into desired, or predetermined, subsets by a variety of techniques, as mentioned above, e.g. FACS using labeled antibody probes to appropriate markers. The objective of the step is to enrich, or preferably isolate, a pure subset, of lymphocytes from the accessible tissue in order to minimize miss-calling subset members from clonotypes identified in the inaccessible tissue. The degree of enrichment depends on the separation or sorting technique employed and available markers for the subsets. In some embodiments, the step of sorting produces at least one subset that is enriched in the target lymphocyte so that at least fifty percent of the sorted population comprises the target lymphocyte. In other embodiments, the step of sorting produces at least one subset that is enriched in the target lymphocyte so that at least eighty

percent of the sorted population comprises the target lymphocyte. In other embodiments, the step of sorting produces at least one subset that is enriched in the target lymphocyte so that at least ninety percent of the sorted population comprises the target lymphocyte. In still other embodiments, for example, when a target lymphocyte belongs to a rare cell population or no efficient probe is available, then the step of sorting may produce a subset that is enriched only to a level of five percent of the sorted population. In such cases, further enrichment may be obtained by using multiple sorting techniques in tandem, e.g. MACS followed by FACS. In some embodiments, as described below, the step of generating a clonotype profile may be implemented by amplifying recombined nucleic acids from the lymphocytes and sequencing isolated nucleic acids from the resulting amplicon. The step of generating may further include coalescing the resulting sequence reads of the sequencing step into clonotypes. Also, the step of generating may further include forming a database of the resulting clonotype sequences which is amenable to analysis, e.g. application of algorithms for comparing such sequences to clonotype sequences of other clonotype profiles.

[0024] As mentioned above, the invention includes methods for determining a prognosis from a state of lymphocyte infiltration into a solid tumor of a patient, wherein such method comprises the steps of: (a) sorting into one or more subsets a sample of lymphocytes from peripheral blood of the patient; (b) generating clonotype profiles for each of the one or more subsets of lymphocytes from the peripheral blood; (c) generating at least one clonotype profile from at least one sample of the solid tumor; and (d) determining numbers, levels, and/or ratios of lymphocytes of each of the one or more subsets. As used herein, a “prognosis” means a prediction of an outcome based on the number, levels, ratios, and/or distribution of functional subsets of lymphocytes in an inaccessible tissue, such as a solid tumor. Outcomes may be patient survival, degree of amelioration of symptoms, reduction of tumor load, or other surrogate measures of improvement or worsening of a disease condition. In some embodiments, a prognosis may be qualitative in that measurements indicate an improvement or worsening, but not a degree of improvement (e.g. number of additional years of survival, etc.) or degree of worsening. In some embodiments, levels of lymphocytes of functional subsets may be relative values, for example, in comparison with levels or concentrations (average or otherwise) in other tissues of the patient or to average levels or ranges in populations of individuals. In one embodiment, relative levels are in comparison with levels in a patient’s peripheral blood of functional subsets of lymphocytes.

Samples

[0025] In accordance with the invention, lymphocytes from an accessible tissue are separated into subsets, which are analyzed to determine clonotypes which, in turn, are used to determine numbers and/or levels of lymphocytes of the different subsets in less accessible tissues; thus, in most embodiments, at least two kinds of sample are obtained, at least one from an accessible tissue and at least one from an inaccessible tissue. In some embodiments, accessible tissues from which samples are taken include, but are not limited to, peripheral blood, bone marrow, lymph fluid, synovial fluid, or the like. In some embodiments, less accessible, or inaccessible, tissues from which samples are taken are solid tissues, such as solid tumors, inflamed tissues associated with autoimmune disease, and the like. Exemplary solid tumors from which less accessible samples are taken include, but are not limited to, melanoma, colorectal, ovarian, gastric, breast, hepatocellular, urothelial, and the like. Of particular interest, are colorectal tumors and melanomas. Exemplary solid tissues related to autoimmune disease include, but are not limited to, connective tissue, joint connective tissue, muscle tissue, skin, lung tissue, small intestine tissue, colon tissue, and the like. In other embodiments, accessible tissue is peripheral blood and less accessible tissues are any tissue that would cause significant patient discomfort to sample. For example, in such embodiments, less accessible tissues may include bone marrow, lymph fluid, synovial fluid, or the like, as well as solid tissues as disclosed above.

[0026] Clonotype profiles are obtained from samples of immune cells (whether in accessible or less accessible tissues), which are present in a wide variety of tissues. Immune cells of interest include T-cells and/or B-cells. T-cells (T lymphocytes) include, for example, cells that express T cell receptors (TCRs). B-cells (B lymphocytes) include, for example, cells that express B cell receptors (BCRs). T-cells include helper T cells (effector T cells or Th cells), cytotoxic T cells (CTLs), memory T cells, and regulatory T cells, which may be distinguished by cell surface markers. In one aspect a sample of T cells includes at least 1,000 T cells; but more typically, a sample includes at least 10,000 T cells, and more typically, at least 100,000 T cells. In another aspect, a sample includes a number of T cells in the range of from 1000 to 1,000,000 cells. A sample of immune cells may also comprise B cells. B-cells include, for example, plasma B cells, memory B cells, B1 cells, B2 cells, marginal-zone B cells, and follicular B cells. B-cells can express immunoglobulins (also referred to as antibodies or B cell receptors). As above, in one aspect a sample of B cells includes at least 1,000 B cells; but more typically, a sample includes at least 10,000 B cells, and more typically, at least 100,000 B cells. In another aspect, a sample includes a number of B cells in the range of from 1000 to 1,000,000 B cells.

[0027] Samples (sometimes referred to as “tissue samples”) used in the methods of the invention can come from a variety of tissues, including, for example, tumor tissue, blood and blood plasma, lymph fluid, cerebrospinal fluid surrounding the brain and the spinal cord, synovial fluid surrounding bone joints, and the like. In one embodiment, the sample is a blood sample. The blood sample can be about 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, or 5.0 mL. The sample can be a tumor biopsy. The biopsy can be from, for example, a tumor of the brain, liver, lung, heart, colon, kidney, or bone marrow. Any biopsy technique used by those skilled in the art can be used for isolating a sample from a subject. For example, a biopsy can be an open biopsy, in which general anesthesia is used. The biopsy can be a closed biopsy, in which a smaller cut is made than in an open biopsy. The biopsy can be a core or incisional biopsy, in which part of the tissue is removed. The biopsy can be an excisional biopsy, in which attempts to remove an entire lesion are made. The biopsy can be a fine needle aspiration biopsy, in which a sample of tissue or fluid is removed with a needle. In some embodiments, multiple samples may be taken from a solid tumor for the purpose of determining a spatial distribution of lymphocyte subsets within or surrounding a solid tumor. In some embodiments, a number of samples from a solid tumor may be in the range of from 2 to 10; in other embodiments, such range may be from 2 to 20.

[0028] A sample or tissue sample, whether accessible or less accessible, includes nucleic acid, for example, DNA (e.g., genomic DNA) or RNA (e.g., messenger RNA). The nucleic acid can be cell-free DNA or RNA, e.g. extracted from the circulatory system, Vlassov et al, Curr. Mol. Med., 10: 142-165 (2010); Swarup et al, FEBS Lett., 581: 795-799 (2007). In the methods of the invention, the amount of RNA or DNA from a subject that can be analyzed varies widely. For generating a clonotype profile, sufficient nucleic acid is obtained in a sample for a useful representation of an individual’s immune receptor repertoire in the tissue. More particularly, for generating a clonotype profile from genomic DNA at least 1 ng of total DNA from T cells or B cells (i.e. about 300 diploid genome equivalents) is extracted from a sample; in another embodiment, at least 2 ng of total DNA (i.e. about 600 diploid genome equivalents) is extracted from a sample; and in another embodiment, at least 3 ng of total DNA (i.e. about 900 diploid genome equivalents) is extracted from a sample. One of ordinary skill would recognize that as the fraction of lymphocytes in a sample decreases, the foregoing minimal amounts of DNA may be increased in order to generate a clonotype profile containing more than about 1000 independent clonotypes. For generating a clonotype profile from RNA, in one embodiment, a sufficient amount of RNA is extracted so that at least 1000 transcripts are obtained which encode distinct TCRs, BCRs, or fragments thereof. The amount of RNA that corresponds to this limit

varies widely from sample to sample depending on the fraction of lymphocytes in a sample, developmental stage of the lymphocytes, sampling techniques, condition of a tissue, and the like. In one embodiment, at least 100 ng of RNA is extracted from a tissue sample containing B cells and/or T cells for the generating of a clonotype profile; in another one embodiment, at least 500 ng of RNA is extracted from a tissue sample containing B cells and/or T cells for the generating of a clonotype profile. RNA used in methods of the invention may be either total RNA extracted from a tissue sample or polyA RNA extracted directly from a tissue sample or from total RNA extracted from a tissue sample. The above nucleic acid extractions may be carried out using commercially available kits, e.g. from Invitrogen (Carlsbad, CA), Qiagen (San Diego, CA), or like vendors. Guidance for extracting RNA is found in Liedtke et al, PCR Methods and Applications, 4: 185-187 (1994); and like references.

[0029] As discussed more fully below (Definitions), a sample of lymphocytes is sufficiently large so that substantially every T cell or B cell with a distinct clonotype is represented therein, thereby forming a repertoire (as the term is used herein). In one embodiment, a sample is taken that contains with a probability of ninety-nine percent every clonotype of a population present at a frequency of .001 percent or greater. In another embodiment, a sample is taken that contains with a probability of ninety-nine percent every clonotype of a population present at a frequency of .0001 percent or greater. In one embodiment, a sample of B cells or T cells includes at least a half million cells, and in another embodiment such sample includes at least one million cells.

[0030] Whenever a source of material from which a sample is taken is scarce, such as, clinical study samples, or the like, DNA from the material may be amplified by a non-biasing technique, such as whole genome amplification (WGA), multiple displacement amplification (MDA); or like technique, e.g. Hawkins et al, Curr. Opin. Biotech., 13: 65-67 (2002); Dean et al, Genome Research, 11: 1095-1099 (2001); Wang et al, Nucleic Acids Research, 32: e76 (2004); Hosono et al, Genome Research, 13: 954-964 (2003); and the like.

[0031] Blood samples are of particular interest as an accessible sample and may be obtained using conventional techniques, e.g. Innis et al, editors, PCR Protocols (Academic Press, 1990); or the like. For example, white blood cells may be separated from blood samples using convention techniques, e.g. RosetteSep kit (Stem Cell Technologies, Vancouver, Canada). Blood samples may range in volume from 100 μ L to 10 mL; in one aspect, blood sample volumes are in the range of from 100 μ L to 2 mL. DNA and/or RNA may then be extracted from such blood sample using conventional techniques for use in methods of the invention, e.g. DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). Optionally, subsets of white blood cells,

e.g. lymphocytes, may be further isolated using conventional techniques, e.g. fluorescently activated cell sorting (FACS)(Becton Dickinson, San Jose, CA), magnetically activated cell sorting (MACS)(Miltenyi Biotec, Auburn, CA), or the like.

[0032] Since the identifying recombinations are present in the DNA of each individual's adaptive immunity cells as well as their associated RNA transcripts, either RNA or DNA can be sequenced in the methods of the provided invention. A recombined sequence from a T-cell or B-cell encoding a T cell receptor or immunoglobulin molecule, or a portion thereof, is referred to as a clonotype. The DNA or RNA can correspond to sequences from T-cell receptor (TCR) genes or immunoglobulin (Ig) genes that encode antibodies. For example, the DNA and RNA can correspond to sequences encoding α , β , γ , or δ chains of a TCR. In a majority of T-cells, the TCR is a heterodimer consisting of an α -chain and β -chain. The TCR α chain is generated by VJ recombination, and the β chain receptor is generated by V(D)J recombination. For the TCR β chain, in humans there are 48 V segments, 2 D segments, and 13 J segments. Several bases may be deleted and others added (called N and P nucleotides) at each of the two junctions. In a minority of T-cells, the TCRs consist of γ and δ delta chains. The TCR γ chain is generated by VJ recombination, and the TCR δ chain is generated by V(D)J recombination (Kenneth Murphy, Paul Travers, and Mark Walport, *Janeway's Immunology* 7th edition, Garland Science, 2007, which is herein incorporated by reference in its entirety).

[0033] The DNA and RNA analyzed in the methods of the invention can correspond to sequences encoding heavy chain immunoglobulins (IgH) with constant regions (α , δ , ϵ , γ , or μ) or light chain immunoglobulins (IgK or IgL) with constant regions λ or κ . Each antibody has two identical light chains and two identical heavy chains. Each chain is composed of a constant (C) and a variable region. For the heavy chain, the variable region is composed of a variable (V), diversity (D), and joining (J) segments. Several distinct sequences coding for each type of these segments are present in the genome. A specific VDJ recombination event occurs during the development of a B-cell, marking that cell to generate a specific heavy chain. Diversity in the light chain is generated in a similar fashion except that there is no D region so there is only VJ recombination. Somatic mutation often occurs close to the site of the recombination, causing the addition or deletion of several nucleotides, further increasing the diversity of heavy and light chains generated by B-cells. The possible diversity of the antibodies generated by a B-cell is then the product of the different heavy and light chains. The variable regions of the heavy and light chains contribute to form the antigen recognition (or binding) region or site. Added to this

diversity is a process of somatic hypermutation which can occur after a specific response is mounted against some epitope.

[0034] In one aspect, where the number of lymphocytes are determined in a sample, a known amount of unique immune receptor rearranged molecules with a known sequence, i.e. known amounts of one or more internal standards, is added to the cDNA or genomic DNA from a sample of unknown quantity. By counting the relative number of molecules that are obtained for the known added sequence compared to the rest of the sequences of the same sample, one can estimate the number of rearranged immune receptor molecules in the initial cDNA sample. (Such techniques for molecular counting are well-known, e.g. Brenner et al, U.S. patent 7,537,897, which is incorporated herein by reference). Data from sequencing the added unique sequence can be used to distinguish the different possibilities if a real time PCR calibration is being used as well, e.g. as disclosed in Faham and Willis (cited above).

Extraction of Nucleic Acids from Fixed Samples

[0035] Fixed tissue samples (for example, from excised tumor tissue, or the like) from which nucleic acids are extracted in conjunction with the invention are typically chemically fixed tissue sample from a disease-related tissue, such as a solid tumor. Chemical fixatives used to produce fixed tissue samples used in the invention include aldehydes, alcohols, and like reagents. Typically, fixed tissue samples used in the invention are fixed with formaldehyde or glutaraldehyde, and in particular, are provided as formalin fixed paraffin embedded (FFPE) tissue samples. Guidance for nucleic acid extraction techniques for use with the invention is disclosed in the following references, which are incorporated by reference: Dedhia et al, Asian Pacific J. Cancer Prev., 8: 55-59 (2007); Okello et al, Analytical Biochemistry, 400: 110-117 (2010); Berezki et al, Pathology Oncology Research, 13(3): 209-214 (2007); Huijsmans et al, BMC Research Notes, 3: 239 (2010); Wood et al, Nucleic Acids Research, 38(14): e151 (2010); Gilbert et al, PLoSOne, 6: e537 (June 2007); Schweiger et al, PLoSOne, 4(5): e5548 (May 2009). In addition, there are several commercially available kits for carrying out nucleic acid extractions from fixed tissue that may be used with the invention using manufacturer's instructions: AllPrep DNA/RNA FFPE Kit (Qiagen, San Diego, CA); Absolutely RNA FFPE Kit (Agilent, Santa Clara, CA); QuickExtract FFPE DNA Extraction Kit (Epicentre, Madison, WI); RecoverAll Total Nucleic Acid Isolation Kit for FFPE (Ambion, Austin, TX); and the like.

[0036] Briefly, nucleic acid extraction may include the following steps: (i) obtaining fixed sample cut in sections about 20 μ m thick or less and in an amount effective for yielding about 6

ng of amplifiable DNA or about 0.5 to 20 ng reverse transcribable and amplifiable RNA; (ii) optionally de-waxing the fixed sample, e.g. by xylene and ethanol washes, d-Limonene and ethanol treatment, microwave treatment, or the like; (iii) optionally treating for reversing fixative-induced cross-linking of DNA, e.g. incubation at 98°C for 15 minutes, or the like; (iv) digesting non-nucleic acid components of the fixed sample, e.g. proteinase K in a conventional buffer, e.g. Tris-HCl, EDTA, NaCl, detergent, followed by heat denaturation of proteinase K, after which the resulting solution optionally may be used directly to generate a clonotype profile to identify correlated clonotypes; (v) and optionally extracting nucleic acid, e.g. phenol:chloroform extraction followed by ethanol precipitation; silica-column based extraction, e.g. QIAamp DNA micro kit (Qiagen, CA); or the like. For RNA isolation, a further step of RNA-specific extraction may be carried out, e.g. RNase inhibitor treatment, DNase treatment, guanidinium thiocyanate/acid extraction, or the like. Additional optional steps may include treating the extracted nucleic acid sample to remove PCR inhibitors, for example, bovine serum albumin or like reagent may be used for this purpose, e.g. Satoh et al, J. Clin. Microbiol., 36(11): 3423-3425 (1998).

[0037] The amount and quality of extracted nucleic acid may be measured in a variety of ways, including but not limited to, PicoGreen Quantitation Assay (Molecular Probes, Eugene, OR); analysis with a 2100 Bioanalyzer (Agilent, Santa Clara, CA); TBS-380 Mini-Fluorometer (Turner Biosystems, Sunnyvale, CA); or the like. In one aspect, a measure of nucleic acid quality may be obtained by amplifying, e.g. in a multiplex PCR, a set of fragments from internal standard genes which have predetermined sizes, e.g. 100, 200, 300, and 400 basepairs, as disclosed in Van Dongen et al, Leukemia, 17: 2257-2317 (2003). After such amplification, fragments are separated by size and bands are quantified to provide a size distribution that reflects the size distribution of fragments of the extracted nucleic acid.

[0038] Nucleic acids extracted from fixed tissues have a distribution of sizes with a typical average size of about 200 nucleotides or less because of the fixation process. Fragments containing clonotypes have sizes that may be in the range of from 100-400 nucleotides; thus, for DNA as the starting material, to ensure the presence of amplifiable clonotypes in the extracted nucleic acid, the number of genome equivalents in a sample must exceed the desired number clonotypes by a significant amount, e.g. typically by 3-6 fold. A similar consideration must be made for RNA as the starting material. If breaks and/or adducts from fixation are randomly distributed along an extracted sequence, then the probability that a region N basepairs in length (for example, containing a clonotype) does not have a break or adduct may be estimated as follows. If each nucleotide has a probability, p, of containing a break or adduct (e.g. p may be

taken as $1/200$, the inverse of the average fragment size), then an estimate of the probability that an N bp stretch will have no break or adduct, is $(1-p)^N$, e.g. Ross, Introduction to Probability Models, Ninth Edition (Academic Press, 2006). The inverse of this quantity is the factor increase in genome equivalents that must be sampled in order to get (on average) the number of desired amplifiable fragments. For example, if at least 1000 amplifiable clonotypes are desired, then there must be at least 1000 sequences encompassing the clonotypes sequences (for example, greater than 300 basepairs (bp)) that do not have breaks or amplification-inhibiting adducts or cross-linkages. For $N=300$ and $p=1/200$, $(1-p)^N \approx 0.22$, so that if a 6 ng sample was required to give about 1000 genome equivalents of intact DNA from unfixed tissue, then about $(1/.22) \times 6$ ng, or 25-30 ng would be required from fixed tissue. For $N=100$ and $p=1/200$, $(1-p)^N \approx 0.61$, so that if a 6 ng sample was required to give about 1000 genome equivalents of intact DNA from unfixed tissue, then about $(1/.61) \times 6$ ng, or 10 ng would be required from fixed tissue. In one aspect, for determination of correlating clonotypes, a number of amplifiable clonotypes is in the range of 1000 to 10000. Accordingly, for fixed tissue samples comprising about 50-100% lymphocytes, a nucleic acid sample from fixed tissue is obtained in an amount in the range of 10-500 ng. For fixed tissue samples comprising about 1-10% lymphocytes, a nucleic acid sample from fixed tissue is obtained in an amount in the range of 1-50 μ g.

Identifying B Cell Isotypes

[0039] In one embodiment, the invention permits the identification of isotypes of B cells that infiltrate an inaccessible tissue. Isotypes of immunoglobulins produced by B lymphocytes may be determined from clonotypes that are designed to include nucleic acid that encodes a portion of the constant region of an immunoglobulin. Thus, in accordance with one aspect of the invention, clonotypes are constructed from sequence reads of nucleotides encoding immunoglobulin heavy chains (IgHs). Such clonotypes of the invention include a portion of a VDJ encoding region and a portion of its associated constant region (or C region). The isotype is determined from the nucleotide sequence encoding the portion of the C region. In one embodiment, the portion encoding the C region is adjacent to the VDJ encoding region, so that a single contiguous sequence may be amplified by a conventional technique, such as polymerase chain reaction (PCR), such as disclosed in Faham and Willis, U.S. patent publication 2011/0207134, which is incorporated herein by reference. The portion of a clonotype encoding C region is used to identify isotype by the presence of characteristic alleles. In one embodiment between 8 and 100 C-region-encoding nucleotides are included in a clonotype; in another embodiment, between 8 and 20 C-region-encoding nucleotides are included in a clonotype. In one embodiment, such C-

region encoding portions are captured during amplification of IgH-encoding sequences as described more fully below. In such amplifications, one or more C-region primers are positioned so that a number of C-region encoding nucleotides in the above ranges are captured in the resulting amplicons.

[0040] There are five types of mammalian Ig heavy chain denoted by the Greek letters: α , δ , ϵ , γ , and μ . The type of heavy chain present defines the class of antibody; these chains are found in IgA, IgD, IgE, IgG, and IgM antibodies, respectively. Distinct heavy chains differ in size and composition; α and γ contain approximately 450 amino acids, while μ and ϵ have approximately 550 amino acids. Each heavy chain has two regions, the constant region and the variable region. The constant region is identical in all antibodies of the same isotype, but differs in antibodies of different isotypes. Heavy chains γ , α and δ have a constant region composed of three tandem (in a line) Ig domains, and a hinge region for added flexibility; heavy chains μ and ϵ have a constant region composed of four immunoglobulin domains. The variable region of the heavy chain differs in antibodies produced by different B cells, but is the same for all antibodies produced by a single B cell or B cell clone. The variable region of each heavy chain is approximately 110 amino acids long and is composed of a single Ig domain. Nucleotide sequences of human (and other) IgH C regions may be obtained from publicly available databases, such as the International Immunogenetics Information System (IMGT) at <http://www.imgt.org>.

As mentioned above, in some embodiments methods of the invention provide for the generation of clonotypes of immunoglobulins containing isotype information. Such methods may be implemented with the following steps: (a) obtaining a sample of nucleic acids from lymphocytes of an individual, the sample comprising recombined sequences each including at least a portion of a C gene segment of a B cell receptor; (b) generating an amplicon from the recombined sequences, each sequence of the amplicon including a portion of a C gene segment; (c) sequencing the amplicon to generate a profile of clonotypes each comprising at least a portion of a VDJ region of a B cell receptor and at least a portion of a C gene segment. From the latter step the isotype of the sampled B lymphocytes are determined by examining the sequence of the C gene segment of its clonotype. In one embodiment, the C gene segment is from a nucleotide sequence encoding an IgH chain of said B cell receptor. Typically, the C gene segment is at one end of a clonotype and a unique recombined sequence portion, e.g. the VDJ portion, is at the other end of the clonotype. In some embodiments, the unique portions of clonotypes comprise at least a portion of a VDJ region.

Amplification of Nucleic Acid Populations

[0041] Amplicons of target populations of nucleic acids may be generated by a variety of amplification techniques. In one aspect of the invention, multiplex PCR is used to amplify members of a mixture of nucleic acids, particularly mixtures comprising recombinant immune molecules such as T cell receptors, or portions thereof. Guidance for carrying out multiplex PCRs of such immune molecules is found in the following references, which are incorporated by reference: Faham and Willis, U.S. patent publication 2011/0207134; Morley, U.S. patent 5,296,351; Gorski, U.S. patent 5,837,447; Dau, U.S. patent 6,087,096; Von Dongen et al, U.S. patent publication 2006/0234234; European patent publication EP 1544308B1; and the like.

[0042] After amplification of DNA from the genome (or amplification of nucleic acid in the form of cDNA by reverse transcribing RNA), the individual nucleic acid molecules can be isolated, optionally re-amplified, and then sequenced individually. Exemplary amplification protocols may be found in van Dongen et al, *Leukemia*, 17: 2257-2317 (2003) or van Dongen et al, U.S. patent publication 2006/0234234, which is incorporated by reference. Briefly, an exemplary protocol is as follows: Reaction buffer: ABI Buffer II or ABI Gold Buffer (Life Technologies, San Diego, CA); 50 μ L final reaction volume; 100 ng sample DNA; 10 pmol of each primer (subject to adjustments to balance amplification as described below); dNTPs at 200 μ M final concentration; $MgCl_2$ at 1.5 mM final concentration (subject to optimization depending on target sequences and polymerase); Taq polymerase (1-2 U/tube); cycling conditions: preactivation 7 min at 95°C; annealing at 60°C; cycling times: 30s denaturation; 30s annealing; 30s extension. Polymerases that can be used for amplification in the methods of the invention are commercially available and include, for example, Taq polymerase, AccuPrime polymerase, or Pfu. The choice of polymerase to use can be based on whether fidelity or efficiency is preferred.

[0043] Real time PCR, picogreen staining, nanofluidic electrophoresis (e.g. LabChip) or UV absorption measurements can be used in an initial step to judge the functional amount of amplifiable material.

[0044] In one aspect, multiplex amplifications are carried out so that relative amounts of sequences in a starting population are substantially the same as those in the amplified population, or amplicon. That is, multiplex amplifications are carried out with minimal amplification bias among member sequences of a sample population. In one embodiment, such relative amounts are substantially the same if each relative amount in an amplicon is within five fold of its value in the starting sample. In another embodiment, such relative amounts are substantially the same

if each relative amount in an amplicon is within two fold of its value in the starting sample. As discussed more fully below, amplification bias in PCR may be detected and corrected using conventional techniques so that a set of PCR primers may be selected for a predetermined repertoire that provide unbiased amplification of any sample.

[0045] In regard to many repertoires based on TCR or BCR sequences, a multiplex amplification optionally uses all the V segments. The reaction is optimized to attempt to get amplification that maintains the relative abundance of the sequences amplified by different V segment primers. Some of the primers are related, and hence many of the primers may “cross talk,” amplifying templates that are not perfectly matched with it. The conditions are optimized so that each template can be amplified in a similar fashion irrespective of which primer amplified it. In other words if there are two templates, then after 1,000 fold amplification both templates can be amplified approximately 1,000 fold, and it does not matter that for one of the templates half of the amplified products carried a different primer because of the cross talk. In subsequent analysis of the sequencing data the primer sequence is eliminated from the analysis, and hence it does not matter what primer is used in the amplification as long as the templates are amplified equally.

[0046] In one embodiment, amplification bias may be avoided by carrying out a two-stage amplification (as described in Faham and Willis, cited above) wherein a small number of amplification cycles are implemented in a first, or primary, stage using primers having tails non-complementary with the target sequences. The tails include primer binding sites that are added to the ends of the sequences of the primary amplicon so that such sites are used in a second stage amplification using only a single forward primer and a single reverse primer, thereby eliminating a primary cause of amplification bias. Preferably, the primary PCR will have a small enough number of cycles (e.g. 5-10) to minimize the differential amplification by the different primers. The secondary amplification is done with one pair of primers and hence the issue of differential amplification is minimal. One percent of the primary PCR is taken directly to the secondary PCR. Thirty-five cycles (equivalent to ~28 cycles without the 100 fold dilution step) used between the two amplifications were sufficient to show a robust amplification irrespective of whether the breakdown of cycles were: one cycle primary and 34 secondary or 25 primary and 10 secondary. Even though ideally doing only 1 cycle in the primary PCR may decrease the amplification bias, there are other considerations. One aspect of this is representation. This plays a role when the starting input amount is not in excess to the number of reads ultimately obtained. For example, if 1,000,000 reads are obtained and starting with 1,000,000 input molecules then taking only representation from 100,000 molecules to the secondary

amplification would degrade the precision of estimating the relative abundance of the different species in the original sample. The 100 fold dilution between the 2 steps means that the representation is reduced unless the primary PCR amplification generated significantly more than 100 molecules. This indicates that a minimum 8 cycles (256 fold), but more comfortably 10 cycle (~1,000 fold), may be used. The alternative to that is to take more than 1% of the primary PCR into the secondary but because of the high concentration of primer used in the primary PCR, a big dilution factor is can be used to ensure these primers do not interfere in the amplification and worsen the amplification bias between sequences. Another alternative is to add a purification or enzymatic step to eliminate the primers from the primary PCR to allow a smaller dilution of it. In this example, the primary PCR was 10 cycles and the second 25 cycles.

Generating Sequence Reads for Clonotypes

[0047] Any high-throughput technique for sequencing nucleic acids can be used in the method of the invention. Preferably, such technique has a capability of generating in a cost-effective manner a volume of sequence data from which at least 1000 clonotypes can be determined, and preferably, from which at least 10,000 to 1,000,000 clonotypes can be determined. DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary, sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, and SOLiD sequencing. Sequencing of the separated molecules has more recently been demonstrated by sequential or single extension reactions using polymerases or ligases as well as by single or sequential differential hybridizations with libraries of probes. These reactions have been performed on many clonal sequences in parallel including demonstrations in current commercial applications of over 100 million sequences in parallel. These sequencing approaches can thus be used to study the repertoire of T-cell receptor (TCR) and/or B-cell receptor (BCR).

[0048] In one aspect of the invention, high-throughput methods of sequencing are employed that comprise a step of spatially isolating individual molecules on a solid surface where they are sequenced in parallel. Such solid surfaces may include nonporous surfaces (such as in Solexa sequencing, e.g. Bentley et al, Nature, 456: 53-59 (2008) or Complete Genomics sequencing, e.g. Drmanac et al, Science, 327: 78-81 (2010)), arrays of wells, which may include bead- or particle-

bound templates (such as with 454, e.g. Margulies et al, Nature, 437: 376-380 (2005) or Ion Torrent sequencing, U.S. patent publication 2010/0137143 or 2010/0304982), micromachined membranes (such as with SMRT sequencing, e.g. Eid et al, Science, 323: 133-138 (2009)), or bead arrays (as with SOLiD sequencing or polony sequencing, e.g. Kim et al, Science, 316: 1481-1414 (2007)).

[0049] In another aspect, such methods comprise amplifying the isolated molecules either before or after they are spatially isolated on a solid surface. Prior amplification may comprise emulsion-based amplification, such as emulsion PCR, or rolling circle amplification. Of particular interest is Solexa-based sequencing where individual template molecules are spatially isolated on a solid surface, after which they are amplified in parallel by bridge PCR to form separate clonal populations, or clusters, and then sequenced, as described in Bentley et al (cited above) and in manufacturer's instructions (e.g. TruSeq™ Sample Preparation Kit and Data Sheet, Illumina, Inc., San Diego, CA, 2010); and further in the following references: U.S. patents 6,090,592; 6,300,070; 7,115,400; and EP0972081B1; which are incorporated by reference. In one embodiment, individual molecules disposed and amplified on a solid surface form clusters in a density of at least 10^5 clusters per cm^2 ; or in a density of at least 5×10^5 per cm^2 ; or in a density of at least 10^6 clusters per cm^2 . In one embodiment, sequencing chemistries are employed having relatively high error rates. In such embodiments, the average quality scores produced by such chemistries are monotonically declining functions of sequence read lengths. In one embodiment, such decline corresponds to 0.5 percent of sequence reads have at least one error in positions 1-75; 1 percent of sequence reads have at least one error in positions 76-100; and 2 percent of sequence reads have at least one error in positions 101-125.

[0050] In one aspect, a sequence-based clonotype profile of an individual is obtained using the following steps: (a) obtaining a nucleic acid sample from T-cells and/or B-cells of the individual; (b) spatially isolating individual molecules derived from such nucleic acid sample, the individual molecules comprising at least one template generated from a nucleic acid in the sample, which template comprises a somatically rearranged region or a portion thereof, each individual molecule being capable of producing at least one sequence read; (c) sequencing said spatially isolated individual molecules; and (d) determining abundances of different sequences of the nucleic acid molecules from the nucleic acid sample to generate the clonotype profile. In another embodiment, a sequence-based clonotype profile may be generated by the following steps: (a) obtaining a sample from the patient comprising T-cells and/or B-cells; (b) amplifying molecules of nucleic acid from the T-cells and/or B-cells of the sample, the molecules of nucleic acid comprising recombined sequences from T-cell receptor genes or immunoglobulin genes; (c)

sequencing the amplified molecules of nucleic acid to form a clonotype profile; and (d) determining a presence, absence and/or level of the one or more patient-specific clonotypes, including any previously unrecorded phylogenetic clonotypes thereof, as taught by Faham and Willis, U.S. patent publication 2011/0207134, which is incorporated herein by reference.

[0051] In one embodiment, each of the somatically rearranged regions comprises a V region and a J region. In another embodiment, the step of sequencing comprises bidirectionally sequencing each of the spatially isolated individual molecules to produce at least one forward sequence read and at least one reverse sequence read. Further to the latter embodiment, at least one of the forward sequence reads and at least one of the reverse sequence reads have an overlap region such that bases of such overlap region are determined by a reverse complementary relationship between such sequence reads. In still another embodiment, each of the somatically rearranged regions comprise a V region and a J region and the step of sequencing further includes determining a sequence of each of the individual nucleic acid molecules from one or more of its forward sequence reads and at least one reverse sequence read starting from a position in a J region and extending in the direction of its associated V region. In another embodiment, individual molecules comprise nucleic acids selected from the group consisting of complete IgH molecules, incomplete IgH molecules, complete IgK molecules, IgK inactive molecules, TCR β molecules, TCR γ molecules, complete TCR δ molecules, and incomplete TCR δ molecules. In another embodiment, the step of sequencing comprises generating the sequence reads having monotonically decreasing quality scores. Further to the latter embodiment, monotonically decreasing quality scores are such that the sequence reads have error rates no better than the following: 0.2 percent of sequence reads contain at least one error in base positions 1 to 50, 0.2 to 1.0 percent of sequence reads contain at least one error in positions 51-75, 0.5 to 1.5 percent of sequence reads contain at least one error in positions 76-100. In another embodiment, the above method comprises the following steps: (a) obtaining a nucleic acid sample from T-cells and/or B-cells of the individual; (b) spatially isolating individual molecules derived from such nucleic acid sample, the individual molecules comprising nested sets of templates each generated from a nucleic acid in the sample and each containing a somatically rearranged region or a portion thereof, each nested set being capable of producing a plurality of sequence reads each extending in the same direction and each starting from a different position on the nucleic acid from which the nested set was generated; (c) sequencing said spatially isolated individual molecules; and (d) determining abundances of different sequences of the nucleic acid molecules from the nucleic acid sample to generate the clonotype profile. In one embodiment, the step of sequencing includes producing a plurality of sequence reads for each of

the nested sets. In another embodiment, each of the somatically rearranged regions comprise a V region and a J region, and each of the plurality of sequence reads starts from a different position in the V region and extends in the direction of its associated J region.

[0052] In one aspect, for each sample from an individual, the sequencing technique used in the methods of the invention generates sequences of least 1000 clonotypes per run; in another aspect, such technique generates sequences of at least 10,000 clonotypes per run; in another aspect, such technique generates sequences of at least 100,000 clonotypes per run; in another aspect, such technique generates sequences of at least 500,000 clonotypes per run; and in another aspect, such technique generates sequences of at least 1,000,000 clonotypes per run. In still another aspect, such technique generates sequences of between 100,000 to 1,000,000 clonotypes per run per individual sample.

[0053] The sequencing technique used in the methods of the provided invention can generate about 30 bp, about 40 bp, about 50 bp, about 60 bp, about 70 bp, about 80 bp, about 90 bp, about 100 bp, about 110, about 120 bp per read, about 150 bp, about 200 bp, about 250 bp, about 300 bp, about 350 bp, about 400 bp, about 450 bp, about 500 bp, about 550 bp, or about 600 bp per read.

Clonotype Determination from Sequence Data

[0054] Constructing clonotypes from sequence read data depends in part on the sequencing method used to generate such data, as the different methods have different expected read lengths and data quality. In one approach, a Solexa sequencer is employed to generate sequence read data for analysis. In one embodiment, a sample is obtained that provides at least $0.5-1.0 \times 10^6$ lymphocytes to produce at least 1 million template molecules, which after optional amplification may produce a corresponding one million or more clonal populations of template molecules (or clusters). For most high throughput sequencing approaches, including the Solexa approach, such over sampling at the cluster level is desirable so that each template sequence is determined with a large degree of redundancy to increase the accuracy of sequence determination. For Solexa-based implementations, preferably the sequence of each independent template is determined 10 times or more. For other sequencing approaches with different expected read lengths and data quality, different levels of redundancy may be used for comparable accuracy of sequence determination. Those of ordinary skill in the art recognize that the above parameters, e.g. sample size, redundancy, and the like, are design choices related to particular applications.

[0055] In one aspect of the invention, sequences of clonotypes (including but not limited to those derived from IgH, TCR α , TCR β , TCR γ , TCR δ , and/or IgL κ (IgK)) may be determined by combining information from one or more sequence reads, for example, along the V(D)J regions of the selected chains. In another aspect, sequences of clonotypes are determined by combining information from a plurality of sequence reads. Such pluralities of sequence reads may include one or more sequence reads along a sense strand (i.e. “forward” sequence reads) and one or more sequence reads along its complementary strand (i.e. “reverse” sequence reads). When multiple sequence reads are generated along the same strand, separate templates are first generated by amplifying sample molecules with primers selected for the different positions of the sequence reads. This concept is illustrated in Fig. 4A where primers (404, 406 and 408) are employed to generate amplicons (410, 412, and 414, respectively) in a single reaction. Such amplifications may be carried out in the same reaction or in separate reactions. In one aspect, whenever PCR is employed, separate amplification reactions are used for generating the separate templates which, in turn, are combined and used to generate multiple sequence reads along the same strand. This latter approach is preferable for avoiding the need to balance primer concentrations (and/or other reaction parameters) to ensure equal amplification of the multiple templates (sometimes referred to herein as “balanced amplification” or “unbias amplification”). The generation of templates in separate reactions is illustrated in Figs. 4B-4C. There a sample containing IgH (400) is divided into three portions (470, 472, and 474) which are added to separate PCRs using J region primers (401) and V region primers (404, 406, and 408, respectively) to produce amplicons (420, 422 and 424, respectively). The latter amplicons are then combined (478) in secondary PCR (480) using P5 and P7 primers to prepare the templates (482) for bridge PCR and sequencing on an Illumina GA sequencer, or like instrument.

[0056] Sequence reads of the invention may have a wide variety of lengths, depending in part on the sequencing technique being employed. For example, for some techniques, several trade-offs may arise in its implementation, for example, (i) the number and lengths of sequence reads per template and (ii) the cost and duration of a sequencing operation. In one embodiment, sequence reads are in the range of from 20 to 400 nucleotides; in another embodiment, sequence reads are in a range of from 30 to 200 nucleotides; in still another embodiment, sequence reads are in the range of from 30 to 120 nucleotides. In one embodiment, 1 to 4 sequence reads are generated for determining the sequence of each clonotype; in another embodiment, 2 to 4 sequence reads are generated for determining the sequence of each clonotype; and in another embodiment, 2 to 3 sequence reads are generated for determining the sequence of each clonotype. In the foregoing embodiments, the numbers given are exclusive of sequence reads used to identify samples from

different individuals. The lengths of the various sequence reads used in the embodiments described below may also vary based on the information that is sought to be captured by the read; for example, the starting location and length of a sequence read may be designed to provide the length of an NDN region as well as its nucleotide sequence; thus, sequence reads spanning the entire NDN region are selected. In other aspects, one or more sequence reads that in combination (but not separately) encompass a D and /or NDN region are sufficient.

[0057] As mentioned above, a variety of algorithms may be used to convert sequence reads into clonotypes. In one embodiment, sequences of clonotypes are determined in part by aligning sequence reads to one or more V region reference sequences and one or more J region reference sequences, and in part by base determination without alignment to reference sequences, such as in the highly variable NDN region. A variety of alignment algorithms may be applied to the sequence reads and reference sequences. For example, guidance for selecting alignment methods is available in Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005), which is incorporated by reference. In one aspect, whenever V reads or C reads (as mentioned above) are aligned to V and J region reference sequences, a tree search algorithm is employed, e.g. as described generally in Gusfield (cited above) and Cormen et al, Introduction to Algorithms, Third Edition (The MIT Press, 2009).

[0058] In another embodiment, an end of at least one forward read and an end of at least one reverse read overlap in an overlap region (e.g. 308 in Fig. 3A), so that the bases of the reads are in a reverse complementary relationship with one another. Thus, for example, if a forward read in the overlap region is “5’-acgttgc”, then a reverse read in a reverse complementary relationship is “5’-gcaacgt” within the same overlap region. In one aspect, bases within such an overlap region are determined, at least in part, from such a reverse complementary relationship. That is, a likelihood of a base call (or a related quality score) in a prospective overlap region is increased if it preserves, or is consistent with, a reverse complementary relationship between the two sequence reads. In one aspect, clonotypes of TCR β and IgH chains (illustrated in Fig. 3A) are determined by at least one sequence read starting in its J region and extending in the direction of its associated V region (referred to herein as a “C read” (304)) and at least one sequence read starting in its V region and extending in the direction of its associated J region (referred to herein as a “V read” (306)). Overlap region (308) may or may not encompass the NDN region (315) as shown in Fig. 3A. Overlap region (308) may be entirely in the J region, entirely in the NDN region, entirely in the V region, or it may encompass a J region-NDN region boundary or a V region-NDN region boundary, or both such boundaries (as illustrated in Fig. 3A). Typically, such sequence reads are generated by extending sequencing primers, e.g. (302) and (310) in Fig.

3A, with a polymerase in a sequencing-by-synthesis reaction, e.g. Metzger, Nature Reviews Genetics, 11: 31-46 (2010); Fuller et al, Nature Biotechnology, 27: 1013-1023 (2009). The binding sites for primers (302) and (310) are predetermined, so that they can provide a starting point or anchoring point for initial alignment and analysis of the sequence reads. In one embodiment, a C read is positioned so that it encompasses the D and/or NDN region of the TCR β or IgH chain and includes a portion of the adjacent V region, e.g. as illustrated in Figs. 3A and 3B. In one aspect, the overlap of the V read and the C read in the V region is used to align the reads with one another. In other embodiments, such alignment of sequence reads is not necessary, e.g. with TCR β chains, so that a V read may only be long enough to identify the particular V region of a clonotype. This latter aspect is illustrated in Fig. 3B. Sequence read (330) is used to identify a V region, with or without overlapping another sequence read, and another sequence read (332) traverses the NDN region and is used to determine the sequence thereof. Portion (334) of sequence read (332) that extends into the V region is used to associate the sequence information of sequence read (332) with that of sequence read (330) to determine a clonotype. For some sequencing methods, such as base-by-base approaches like the Solexa sequencing method, sequencing run time and reagent costs are reduced by minimizing the number of sequencing cycles in an analysis. Optionally, as illustrated in Fig. 3A, amplicon (300) is produced with sample tag (312) to distinguish between clonotypes originating from different biological samples, e.g. different patients. Sample tag (312) may be identified by annealing a primer to primer binding region (316) and extending it (314) to produce a sequence read across tag (312), from which sample tag (312) is decoded.

[0059] The IgH chain is more challenging to analyze than TCR β chain because of at least two factors: i) the presence of somatic mutations makes the mapping or alignment more difficult, and ii) the NDN region is larger so that it is often not possible to map a portion of the V segment to the C read. In one aspect of the invention, this problem is overcome by using a plurality of primer sets for generating V reads, which are located at different locations along the V region, preferably so that the primer binding sites are nonoverlapping and spaced apart, and with at least one primer binding site adjacent to the NDN region, e.g. in one embodiment from 5 to 50 bases from the V-NDN junction, or in another embodiment from 10 to 50 bases from the V-NDN junction. The redundancy of a plurality of primer sets minimizes the risk of failing to detect a clonotype due to a failure of one or two primers having binding sites affected by somatic mutations. In addition, the presence of at least one primer binding site adjacent to the NDN region makes it more likely that a V read will overlap with the C read and hence effectively extend the length of the C read. This allows for the generation of a continuous sequence that

spans all sizes of NDN regions and that can also map substantially the entire V and J regions on both sides of the NDN region. Embodiments for carrying out such a scheme are illustrated in Figs. 4A and 4D. In Fig. 4A, a sample comprising IgH chains (400) are sequenced by generating a plurality amplicons for each chain by amplifying the chains with a single set of J region primers (401) and a plurality (three shown) of sets of V region (402) primers (404, 406, 408) to produce a plurality of nested amplicons (e.g., 410, 412, 414) all comprising the same NDN region and having different lengths encompassing successively larger portions (411, 413, 415) of V region (402). Members of a nested set may be grouped together after sequencing by noting the identify (or substantial identity) of their respective NDN, J and/or C regions, thereby allowing reconstruction of a longer V(D)J segment than would be the case otherwise for a sequencing platform with limited read length and/or sequence quality. In one embodiment, the plurality of primer sets may be a number in the range of from 2 to 5. In another embodiment the plurality is 2-3; and still another embodiment the plurality is 3. The concentrations and positions of the primers in a plurality may vary widely. Concentrations of the V region primers may or may not be the same. In one embodiment, the primer closest to the NDN region has a higher concentration than the other primers of the plurality, e.g. to insure that amplicons containing the NDN region are represented in the resulting amplicon. In a particular embodiment where a plurality of three primers is employed, a concentration ratio of 60:20:20 is used. One or more primers (e.g. 435 and 437 in Fig. 4D) adjacent to the NDN region (444) may be used to generate one or more sequence reads (e.g. 434 and 436) that overlap the sequence read (442) generated by J region primer (432), thereby improving the quality of base calls in overlap region (440). Sequence reads from the plurality of primers may or may not overlap the adjacent downstream primer binding site and/or adjacent downstream sequence read. In one embodiment, sequence reads proximal to the NDN region (e.g. 436 and 438) may be used to identify the particular V region associated with the clonotype. Such a plurality of primers reduces the likelihood of incomplete or failed amplification in case one of the primer binding sites is hypermutated during immunoglobulin development. It also increases the likelihood that diversity introduced by hypermutation of the V region will be capture in a clonotype sequence. A secondary PCR may be performed to prepare the nested amplicons for sequencing, e.g. by amplifying with the P5 (401) and P7 (404, 406, 408) primers as illustrated to produce amplicons (420, 422, and 424), which may be distributed as single molecules on a solid surface, where they are further amplified by bridge PCR, or like technique.

[0060] Base calling in NDN regions (particularly of IgH chains) can be improved by using the codon structure of the flanking J and V regions, as illustrated in Fig. 4E. (As used herein,

“codon structure” means the codons of the natural reading frame of segments of TCR or BCR transcripts or genes outside of the NDN regions, e.g. the V region, J region, or the like.) There amplicon (450), which is an enlarged view of the amplicon of Fig. 4B, is shown along with the relative positions of C read (442) and adjacent V read (434) above and the codon structures (452 and 454) of V region (430) and J region (446), respectively, below. In accordance with this aspect of the invention, after the codon structures (452 and 454) are identified by conventional alignment to the V and J reference sequences, bases in NDN region (456) are called (or identified) one base at a time moving from J region (446) toward V region (430) and in the opposite direction from V region (430) toward J region (446) using sequence reads (434) and (442). Under normal biological conditions, only the recombined TCR or IgH sequences that have in frame codons from the V region through the NDN region and to the J region are expressed as proteins. That is, of the variants generated somatically only ones expressed are those whose J region and V region codon frames are in-frame with one another and remain in-frame through the NDN region. (Here the correct frames of the V and J regions are determined from reference sequences). If an out-of-frame sequence is identified based one or more low quality base calls, the corresponding clonotype is flagged for re-evaluation or as a potential disease-related anomaly. If the sequence identified is in-frame and based on high quality base calls, then there is greater confidence that the corresponding clonotype has been correctly called. Accordingly, in one aspect, the invention includes a method of determining V(D)J-based clonotypes from bidirectional sequence reads comprising the steps of: (a) generating at least one J region sequence read that begins in a J region and extends into an NDN region and at least one V region sequence read that begins in the V regions and extends toward the NDN region such that the J region sequence read and the V region sequence read are overlapping in an overlap region, and the J region and the V region each have a codon structure; (b) determining whether the codon structure of the J region extended into the NDN region is in frame with the codon structure of the V region extended toward the NDN region. In a further embodiment, the step of generating includes generating at least one V region sequence read that begins in the V region and extends through the NDN region to the J region, such that the J region sequence read and the V region sequence read are overlapping in an overlap region.

[0061] In some embodiments, IgH-based clonotypes that have undergone somatic hypermutation may be determined as follows. A somatic mutation is defined as a sequenced base that is different from the corresponding base of a reference sequence (of the relevant segment, usually V, J or C) and that is present in a statistically significant number of reads. In one embodiment, C reads may be used to find somatic mutations with respect to the mapped J segment and likewise

V reads for the V segment. Only pieces of the C and V reads are used that are either directly mapped to J or V segments or that are inside the clonotype extension up to the NDN boundary. In this way, the NDN region is avoided and the same 'sequence information' is not used for mutation finding that was previously used for clonotype determination (to avoid erroneously classifying as mutations nucleotides that are really just different recombined NDN regions). For each segment type, the mapped segment (major allele) is used as a scaffold and all reads are considered which have mapped to this allele during the read mapping phase. Each position of the reference sequences where at least one read has mapped is analyzed for somatic mutations. In one embodiment, the criteria for accepting a non-reference base as a valid mutation include the following: 1) at least N reads with the given mutation base, 2) at least a given fraction N/M reads (where M is the total number of mapped reads at this base position) and 3) a statistical cut based on the binomial distribution, the average Q score of the N reads at the mutation base as well as the number (M-N) of reads with a non-mutation base. Preferably, the above parameters are selected so that the false discovery rate of mutations per clonotype is less than 1 in 1000, and more preferably, less than 1 in 10000.

[0062] Sequence-tag-based methods are an alternative to the above approaches for constructing clonotypes from sequence data. Sequence data typically comprises a large collection of sequence reads, i.e. sequences of base calls and associated quality scores, from a DNA sequencer used to analyze the immune molecules. A key challenge in constructing clonotype profiles is to rapidly and accurately distinguish sequence reads that contain genuine differences from those that contain errors from non-biological sources, such as the extraction steps, sequencing chemistry, amplification chemistry, or the like. In one approach to generating clonotypes, a unique sequence tag may be attached to each clonotype in a sample to assist in determining whether sequence reads of such conjugates are derived from the same original clonotype before amplification or sequencing. Sequence tags may be attached to the somatically recombined nucleic acid molecules to form tag-molecule conjugates wherein each recombined nucleic acid of such a conjugate has a unique sequence tag. Usually such attachment is made after nucleic acid molecules are extracted from a sample containing T cells and/or B cells. Preferably, such unique sequence tags differ greatly from one another as determined by conventional distance measures for sequences, such as, Hamming distance, or the like; thus, copies of each sequence tag in tag-molecule conjugates remains far closer to its ancestral tag sequence than to that of any other unique tag sequence, even with a high rate of sequencing or amplification errors introduced by steps of the invention. For example, if 16-mer sequence tags are employed and each such tag on a set of clonotypes has a Hamming distance of at least fifty percent, of eight

nucleotides, from every other sequence tag on the clonotypes, then at least eight sequencing or amplification errors would be necessary to transform one such tag into another for a mis-read of a sequence tag (and the incorrect grouping of a sequence read of a clonotype with the wrong sequence tag). In one embodiment, sequence tags are selected so that after attachment to recombinant nucleic acids molecules to form tag-molecule conjugates, the Hamming distance between tags of the tag-molecule conjugates is a number at least twenty-five percent of the total length of such sequence tags (that is, each sequence tag differs in sequence from every other such tag in at least 25 percent of its nucleotides); in another embodiment, the Hamming distance between such sequence tags is a number at least 50 percent of the total length of such sequence tags.

[0063] In one aspect, the above approach is implemented by the following steps: (a) obtaining a sample from an individual comprising T-cells and/or B-cells; (b) attaching sequence tags to molecules of recombinant nucleic acids of T-cell receptor genes or immunoglobulin genes of the T-cells and/or B-cells to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; and (e) aligning sequence reads of like sequence tags to determine sequence reads corresponding to the same clonotypes of the repertoire. Samples containing B-cells or T-cells are obtained using conventional techniques, as described more fully below. In the step of attaching sequence tags, preferably sequence tags are not only unique but also are sufficiently different from one another that the likelihood of even a large number of sequencing or amplification errors transforming one sequence tag into another would be close to zero. After attaching sequence tags, amplification of the tag-molecule conjugate is necessary for most sequencing technologies; however, whenever single-molecule sequencing technologies are employed an amplification step is optional. Single molecule sequencing technologies include, but are not limited to, single molecule real-time (SMRT) sequencing, nanopore sequencing, or the like, e.g. U.S. patents 7,313,308; 8,153,375; 7,907,800; 7,960,116; 8,137,569; Manrao et al, Nature Biotechnology, 4(8): 2685-2693 (2012); and the like.

[0064] In another aspect, the invention includes a method for determining the number of lymphocytes in a sample by counting unique sequence tags. Even without sequence tags, clonotypes of TCR β or IgH genes, particularly those including the V(D)J regions, provide for a lymphocyte and its clones a unique marker. Whenever recombinant nucleic acids are obtained from genomic DNA, then a count of lymphocytes in a sample may be estimated by the number of unique clonotypes that are counted after sequencing. This approach breaks down whenever there are significant clonal populations of identical lymphocytes associated with the same

clonotype. The use of sequence tags overcomes this short coming and is especially useful for providing counts of lymphocytes in patients suffering from many lymphoid disorders, such as lymphomas or leukemias. In accordance with one aspect of the invention, sequence tags may be used to obtain an absolute count of lymphocytes in a sample regardless of whether there is a large dominant clone present, such as with leukemia. Such a method may be implemented with the steps: (a) obtaining a sample from an individual comprising lymphocytes; (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or of immunoglobulin genes of the lymphocytes to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; and (e) counting the number of distinct sequence tags to determine the number of lymphocytes in the sample.

[0065] In some embodiments, sequence tags are attached to recombined nucleic acid molecules of a sample by labeling by sampling, e.g. as disclosed by Brenner et al, U.S. patent 5,846,719; Brenner et al, U.S. patent 7,537,897; Macevicz, International patent publication WO 2005/111242; and the like, which are incorporated herein by reference. In labeling by sampling, polynucleotides of a population to be labeled (or uniquely tagged) are used to sample (by attachment, linking, or the like) sequence tags of a much larger population. That is, if the population of polynucleotides has K members (including replicates of the same polynucleotide) and the population of sequence tags has N members, then $N \gg K$. In one embodiment, the size of a population of sequence tags used with the invention is at least 10 times the size of the population of clonotypes in a sample; in another embodiment, the size of a population of sequence tags used with the invention is at least 100 times the size of the population of clonotypes in a sample; and in another embodiment, the size of a population of sequence tags used with the invention is at least 1000 times the size of the population of clonotypes in a sample. In other embodiments, a size of sequence tag population is selected so that substantially every clonotype in a sample will have a unique sequence tag whenever such clonotypes are combined with such sequence tag population, e.g. in an attachment reaction, such as a ligation reaction, amplification reaction, or the like. In some embodiments, substantially every clonotype means at least 90 percent of such clonotypes will have a unique sequence tag; in other embodiments, substantially every clonotype means at least 99 percent of such clonotypes will have a unique sequence tag; in other embodiments, substantially every clonotype means at least 99.9 percent of such clonotypes will have a unique sequence tag. In many tissue samples or biopsies the number of T cells or B cells may be up to or about 1 million cells; thus, in some

embodiments of the invention employing such samples, the number of unique sequence tags employed in labeling by sampling is at least 10^8 or in other embodiments at least 10^9 .

[0066] In such embodiments, in which up to 1 million clonotypes are labeled by sampling, large sets of sequence tags may be efficiently produced by combinatorial synthesis by reacting a mixture of all four nucleotide precursors at each addition step of a synthesis reaction, e.g. as disclosed in Church, U.S. patent 5,149,625, which is incorporated by reference. The result is a set of sequence tags having a structure of " $N_1N_2 \dots N_k$ " where each $N_i=A, C, G$ or T and k is the number of nucleotides in the tags. The number of sequence tags in a set of sequence tags made by such combinatorial synthesis is 4^k . Thus, a set of such sequence tags with k at least 14, or k in the range of about 14 to 18, is appropriate for attaching sequence tags to a 10^6 -member population of molecules by labeling by sampling.

[0067] A variety of different attachment reactions may be used to attach unique tags to substantially every clonotype in a sample. In one embodiment, such attachment is accomplished by combining a sample containing recombined nucleic acid molecules (which, in turn, comprise clonotype sequences) with a population or library of sequence tags so that members of the two populations of molecules can randomly combine and become associated or linked, e.g. covalently. In such tag attachment reactions, clonotype sequences comprise linear single or double stranded polynucleotides and sequence tags are carried by reagents such as amplification primers, such as PCR primers, ligation adaptors, circularizable probes, plasmids, or the like. Several such reagents capable of carrying sequence tag populations are disclosed in Macevicz, U.S. patent 8,137,936; Faham et al, U.S. patent 7,862,999; Landegren et al, U.S. patent 8,053,188; Unrau and Deugau, Gene, 145: 163-169 (1994); Church, U.S. patent 5,149,625; and the like, which are incorporated herein by reference.

TCR β Repertoire Analysis

[0068] In this example, TCR β chains are analyzed. The analysis includes amplification, sequencing, and analyzing the TCR β sequences. One primer is complementary to a common sequence in C β 1 and C β 2, and there are 34 V primers capable of amplifying all 48 V segments. C β 1 or C β 2 differ from each other at position 10 and 14 from the J/C junction. The primer for C β 1 and C β 2 ends at position 16 bp and has no preference for C β 1 or C β 2. The 34 V primers are modified from an original set of primers disclosed in Van Dongen et al, U.S. patent publication 2006/0234234, which is incorporated herein by reference. The modified primers are disclosed in

Faham et al, U.S. patent publication 2010/0151471, which is also incorporated herein by reference.

[0069] The Illumina Genome Analyzer is used to sequence the amplicon produced by the above primers. A two-stage amplification is performed on messenger RNA transcripts (200), as illustrated in Figs. 2A-2B, the first stage employing the above primers and a second stage to add common primers for bridge amplification and sequencing. As shown in FIG. 2A, a primary PCR is performed using on one side a 20 bp primer (202) whose 3' end is 16 bases from the J/C junction (204) and which is perfectly complementary to C β 1(203) and the two alleles of C β 2. In the V region (206) of RNA transcripts (200), primer set (212) is provided which contains primer sequences complementary to the different V region sequences (34 in one embodiment). Primers of set (212) also contain a non-complementary tail (214) that produces amplicon (216) having primer binding site (218) specific for P7 primers (220). After a conventional multiplex PCR, amplicon (216) is formed that contains the highly diverse portion of the J(D)V region (206, 208, and 210) of the mRNA transcripts and common primer binding sites (203 and 218) for a secondary amplification to add a sample tag (221) and primers (220 and 222) for cluster formation by bridge PCR. In the secondary PCR, on the same side of the template, a primer (222 in Fig. 2B and referred to herein as "C10-17-P5") is used that has at its 3' end the sequence of the 10 bases closest to the J/C junction, followed by 17 bp with the sequence of positions 15-31 from the J/C junction, followed by the P5 sequence (224), which plays a role in cluster formation by bridge PCR in Solexa sequencing. (When the C10-17-P5 primer (222) anneals to the template generated from the first PCR, a 4 bp loop (position 11-14) is created in the template, as the primer hybridizes to the sequence of the 10 bases closest to the J/C junction and bases at positions 15-31 from the J/C junction. The looping of positions 11-14 eliminates differential amplification of templates carrying C β 1 or C β 2. Sequencing is then done with a primer complementary to the sequence of the 10 bases closest to the J/C junction and bases at positions 15-31 from the J/C junction (this primer is called C'). C10-17-P5 primer can be HPLC purified in order to ensure that all the amplified material has intact ends that can be efficiently utilized in the cluster formation.)

[0070] In FIG. 2A, the length of the overhang on the V primers (212) is preferably 14 bp. The primary PCR is helped with a shorter overhang (214). Alternatively, for the sake of the secondary PCR, the overhang in the V primer is used in the primary PCR as long as possible because the secondary PCR is priming from this sequence. A minimum size of overhang (214) that supports an efficient secondary PCR was investigated. Two series of V primers (for two different V segments) with overhang sizes from 10 to 30 with 2 bp steps were made. Using the

appropriate synthetic sequences, the first PCR was performed with each of the primers in the series and gel electrophoresis was performed to show that all amplified.

[0071] As illustrated in FIG. 2A, the primary PCR uses 34 different V primers (212) that anneal to V region (206) of RNA templates (200) and contain a common 14 bp overhang on the 5' tail. The 14 bp is the partial sequence of one of the Illumina sequencing primers (termed the Read 2 primer). The secondary amplification primer (220) on the same side includes P7 sequence, a tag (221), and Read 2 primer sequence (223) (this primer is called Read2_tagX_P7). The P7 sequence is used for cluster formation. Read 2 primer and its complement are used for sequencing the V segment and the tag respectively. A set of 96 of these primers with tags numbered 1 through 96 are created (see below). These primers are HPLC purified in order to ensure that all the amplified material has intact ends that can be efficiently utilized in the cluster formation.

[0072] As mentioned above, the second stage primer, C-10-17-P5 (222, FIG. 2B) has interrupted homology to the template generated in the first stage PCR. The efficiency of amplification using this primer has been validated. An alternative primer to C-10-17-P5, termed CsegP5, has perfect homology to the first stage C primer and a 5' tail carrying P5. The efficiency of using C-10-17-P5 and CsegP5 in amplifying first stage PCR templates was compared by performing real time PCR. In several replicates, it was found that PCR using the C-10-17-P5 primer had little or no difference in efficiency compared with PCR using the CsegP5 primer.

[0073] Amplicon (230) resulting from the 2-stage amplification illustrated in Figs. 2A-2C has the structure typically used with the Illumina sequencer as shown in FIG. 2C. Two primers that anneal to the outmost part of the molecule, Illumina primers P5 and P7 are used for solid phase amplification of the molecule (cluster formation). Three sequence reads are done per molecule. The first read of 100 bp is done with the C' primer, which has a melting temperature that is appropriate for the Illumina sequencing process. The second read is 6 bp long only and is solely for the purpose of identifying the sample tag. It is generated using a tag primer provided by the manufacturer (Illumina). The final read is the Read 2 primer, also provided by the manufacturer (Illumina). Using this primer, a 100 bp read in the V segment is generated starting with the 1st PCR V primer sequence.

[0074] While the present invention has been described with reference to several particular example embodiments, those skilled in the art will recognize that many changes may be made thereto without departing from the spirit and scope of the present invention. The present

invention is applicable to a variety of sensor implementations and other subject matter, in addition to those discussed above.

Definitions

[0075] Unless otherwise specifically defined herein, terms and symbols of nucleic acid chemistry, biochemistry, genetics, and molecular biology used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, DNA Replication, Second Edition (W.H. Freeman, New York, 1992); Lehninger, Biochemistry, Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human Molecular Genetics, Second Edition (Wiley-Liss, New York, 1999); Abbas et al, Cellular and Molecular Immunology, 6th edition (Saunders, 2007).

[0076] “Aligning” means a method of comparing a test sequence, such as a sequence read, to one or more reference sequences to determine which reference sequence or which portion of a reference sequence is closest based on some sequence distance measure. An exemplary method of aligning nucleotide sequences is the Smith Waterman algorithm. Distance measures may include Hamming distance, Levenshtein distance, or the like. Distance measures may include a component related to the quality values of nucleotides of the sequences being compared.

[0077] “Amplicon” means the product of a polynucleotide amplification reaction; that is, a clonal population of polynucleotides, which may be single stranded or double stranded, which are replicated from one or more starting sequences. The one or more starting sequences may be one or more copies of the same sequence, or they may be a mixture of different sequences. Preferably, amplicons are formed by the amplification of a single starting sequence. Amplicons may be produced by a variety of amplification reactions whose products comprise replicates of the one or more starting, or target, nucleic acids. In one aspect, amplification reactions producing amplicons are “template-driven” in that base pairing of reactants, either nucleotides or oligonucleotides, have complements in a template polynucleotide that are required for the creation of reaction products. In one aspect, template-driven reactions are primer extensions with a nucleic acid polymerase or oligonucleotide ligations with a nucleic acid ligase. Such reactions include, but are not limited to, polymerase chain reactions (PCRs), linear polymerase reactions, nucleic acid sequence-based amplification (NASBAs), rolling circle amplifications, and the like, disclosed in the following references that are incorporated herein by reference: Mullis et al, U.S. patents 4,683,195; 4,965,188; 4,683,202; 4,800,159 (PCR); Gelfand et al, U.S. patent 5,210,015 (real-time PCR with “taqman” probes); Wittwer et al, U.S. patent 6,174,670; Kacian et al, U.S. patent 5,399,491 (“NASBA”); Lizardi, U.S. patent 5,854,033; Aono et al,

Japanese patent publ. JP 4-262799 (rolling circle amplification); and the like. In one aspect, amplicons of the invention are produced by PCRs. An amplification reaction may be a “real-time” amplification if a detection chemistry is available that permits a reaction product to be measured as the amplification reaction progresses, e.g. “real-time PCR” described below, or “real-time NASBA” as described in Leone et al, *Nucleic Acids Research*, 26: 2150-2155 (1998), and like references. As used herein, the term “amplifying” means performing an amplification reaction. A “reaction mixture” means a solution containing all the necessary reactants for performing a reaction, which may include, but not be limited to, buffering agents to maintain pH at a selected level during a reaction, salts, co-factors, scavengers, and the like.

[0078] “Clonotype” means a recombined nucleotide sequence of a T cell or B cell encoding a T cell receptor (TCR) or B cell receptor (BCR), or a portion thereof. In one aspect, a collection of all the distinct clonotypes of a population of lymphocytes of an individual is a repertoire of such population, e.g. Arstila et al, *Science*, 286: 958-961 (1999); Yassai et al, *Immunogenetics*, 61: 493-502 (2009); Kedzierska et al, *Mol. Immunol.*, 45(3): 607-618 (2008); and the like. As used herein, “clonotype profile,” or “repertoire profile,” is a tabulation of clonotypes of a sample of T cells and/or B cells (such as a peripheral blood sample containing such cells) that includes substantially all of the repertoire’s clonotypes and their relative abundances. “Clonotype profile,” “repertoire profile,” and “repertoire” are used herein interchangeably. (That is, the term “repertoire,” as discussed more fully below, means a repertoire measured from a sample of lymphocytes). In one aspect of the invention, clonotypes comprise portions of an immunoglobulin heavy chain (IgH) or a TCR β chain. In other aspects of the invention, clonotypes may be based on other recombined molecules, such as immunoglobulin light chains or TCR α chains, or portions thereof.

[0079] “Complementarity determining regions” (CDRs) mean regions of an immunoglobulin (i.e., antibody) or T cell receptor where the molecule complements an antigen's conformation, thereby determining the molecule's specificity and contact with a specific antigen. T cell receptors and immunoglobulins each have three CDRs: CDR1 and CDR2 are found in the variable (V) domain, and CDR3 includes some of V, all of diverse (D) (heavy chains only) and joint (J), and some of the constant (C) domains.

[0080] “Percent homologous,” “percent identical,” or like terms used in reference to the comparison of a reference sequence and another sequence (“comparison sequence”) mean that in an optimal alignment between the two sequences, the comparison sequence is identical to the reference sequence in a number of subunit positions equivalent to the indicated percentage, the

subunits being nucleotides for polynucleotide comparisons or amino acids for polypeptide comparisons. As used herein, an "optimal alignment" of sequences being compared is one that maximizes matches between subunits and minimizes the number of gaps employed in constructing an alignment. Percent identities may be determined with commercially available implementations of algorithms, such as that described by Needleman and Wunsch, *J. Mol. Biol.*, 48: 443-453 (1970) ("GAP" program of Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI), or the like. Other software packages in the art for constructing alignments and calculating percentage identity or other measures of similarity include the "BestFit" program, based on the algorithm of Smith and Waterman, *Advances in Applied Mathematics*, 2: 482-489 (1981) (Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI). In other words, for example, to obtain a polynucleotide having a nucleotide sequence at least 95 percent identical to a reference nucleotide sequence, up to five percent of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to five percent of the total number of nucleotides in the reference sequence may be inserted into the reference sequence.

[0081] "Flow system" means any instrument or device (i) that is capable of constraining particles or cells to move in a collinear path in a fluid stream by or through one or more detection stations which collect multiparameter data related to the particles or cells and (ii) that is capable of enumerating or sorting such particles based on the collected multiparameter data. Flow systems have a wide variety of forms and use a wide variety of techniques to achieve such functions, as exemplified by the following references that are incorporated by reference Shapiro, *Practical Flow Cytometry*, Fourth Edition (Wiley-Liss, 2003), Bonner et al, *Rev Sci Instruments*, 43 404 (1972), Huh et al, *Physiol Meas*, 26 R73-98 (2005), Ateya et al, *Anal Bioanal Chem*, 391 1485-1498 (2008), Bohm et al, U S patent 7, 157,274; Wang et al, U S patent 7,068,874, and the like. Flow systems may comprise fluidics systems having components wherein a sample fluid stream is inserted into a sheath fluid stream so that particles or cells in the sample fluid are constrained to move in a collinear path, which may take place in a cuvette, other chamber that serves as a detection station, or in a nozzle or other structure, for creating a stream-in-air jet, which may then be manipulated electrically, e.g. as with fluorescence-activated cell sorting (FACS) instruments. Flow systems, flow cytometers, and flow sorters and common applications thereof are disclosed in one or more of the following references, which are incorporated by reference: Robinson et al (Editors) *Current Protocols in Cytometry* (John Wiley & Sons, 2007); Shapiro, *Practical Flow Cytometry*, Fourth Edition (Wiley-Liss, 2003); Owens et al (Editors), *Flow Cytometry Principles for Clinical Laboratory Practice: Quality Assurance for Quantitative*

Immunophenotyping (Wiley-Liss, 1994); Ormerod (Editor) Flow Cytometry: A Practical Approach (Oxford University Press, 2000); and the like.

[0082] “Polymerase chain reaction,” or “PCR,” means a reaction for the in vitro amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA. In other words, PCR is a reaction for making multiple copies or replicates of a target nucleic acid flanked by primer binding sites, such reaction comprising one or more repetitions of the following steps: (i) denaturing the target nucleic acid, (ii) annealing primers to the primer binding sites, and (iii) extending the primers by a nucleic acid polymerase in the presence of nucleoside triphosphates. Usually, the reaction is cycled through different temperatures optimized for each step in a thermal cycler instrument. Particular temperatures, durations at each step, and rates of change between steps depend on many factors well-known to those of ordinary skill in the art, e.g. exemplified by the references: McPherson et al, editors, PCR: A Practical Approach and PCR2: A Practical Approach (IRL Press, Oxford, 1991 and 1995, respectively). For example, in a conventional PCR using Taq DNA polymerase, a double stranded target nucleic acid may be denatured at a temperature $>90^{\circ}\text{C}$, primers annealed at a temperature in the range $50\text{--}75^{\circ}\text{C}$, and primers extended at a temperature in the range $72\text{--}78^{\circ}\text{C}$. The term “PCR” encompasses derivative forms of the reaction, including but not limited to, RT-PCR, real-time PCR, nested PCR, quantitative PCR, multiplexed PCR, and the like. Reaction volumes range from a few hundred nanoliters, e.g. 200 nL, to a few hundred μL , e.g. 200 μL . “Reverse transcription PCR,” or “RT-PCR,” means a PCR that is preceded by a reverse transcription reaction that converts a target RNA to a complementary single stranded DNA, which is then amplified, e.g. Tecott et al, U.S. patent 5,168,038, which patent is incorporated herein by reference. “Real-time PCR” means a PCR for which the amount of reaction product, i.e. amplicon, is monitored as the reaction proceeds. There are many forms of real-time PCR that differ mainly in the detection chemistries used for monitoring the reaction product, e.g. Gelfand et al, U.S. patent 5,210,015 (“taqman”); Wittwer et al, U.S. patents 6,174,670 and 6,569,627 (intercalating dyes); Tyagi et al, U.S. patent 5,925,517 (molecular beacons); which patents are incorporated herein by reference. Detection chemistries for real-time PCR are reviewed in Mackay et al, Nucleic Acids Research, 30: 1292-1305 (2002), which is also incorporated herein by reference. “Nested PCR” means a two-stage PCR wherein the amplicon of a first PCR becomes the sample for a second PCR using a new set of primers, at least one of which binds to an interior location of the first amplicon. As used herein, “initial primers” in reference to a nested amplification reaction mean the primers used to generate a first amplicon, and “secondary primers” mean the one or more primers used to generate a second, or nested, amplicon.

“Multiplexed PCR” means a PCR wherein multiple target sequences (or a single target sequence and one or more reference sequences) are simultaneously carried out in the same reaction mixture, e.g. Bernard et al, *Anal. Biochem.*, 273: 221-228 (1999)(two-color real-time PCR). Usually, distinct sets of primers are employed for each sequence being amplified. Typically, the number of target sequences in a multiplex PCR is in the range of from 2 to 50, or from 2 to 40, or from 2 to 30. “Quantitative PCR” means a PCR designed to measure the abundance of one or more specific target sequences in a sample or specimen. Quantitative PCR includes both absolute quantitation and relative quantitation of such target sequences. Quantitative measurements are made using one or more reference sequences or internal standards that may be assayed separately or together with a target sequence. The reference sequence may be endogenous or exogenous to a sample or specimen, and in the latter case, may comprise one or more competitor templates. Typical endogenous reference sequences include segments of transcripts of the following genes: β -actin, GAPDH, β_2 -microglobulin, ribosomal RNA, and the like. Techniques for quantitative PCR are well-known to those of ordinary skill in the art, as exemplified in the following references that are incorporated by reference: Freeman et al, *Biotechniques*, 26: 112-126 (1999); Becker-Andre et al, *Nucleic Acids Research*, 17: 9437-9447 (1989); Zimmerman et al, *Biotechniques*, 21: 268-279 (1996); Diviacco et al, *Gene*, 122: 3013-3020 (1992); Becker-Andre et al, *Nucleic Acids Research*, 17: 9437-9446 (1989); and the like.

[0083] “Primer” means an oligonucleotide, either natural or synthetic that is capable, upon forming a duplex with a polynucleotide template, of acting as a point of initiation of nucleic acid synthesis and being extended from its 3' end along the template so that an extended duplex is formed. Extension of a primer is usually carried out with a nucleic acid polymerase, such as a DNA or RNA polymerase. The sequence of nucleotides added in the extension process is determined by the sequence of the template polynucleotide. Usually primers are extended by a DNA polymerase. Primers usually have a length in the range of from 14 to 40 nucleotides, or in the range of from 18 to 36 nucleotides. Primers are employed in a variety of nucleic acid amplification reactions, for example, linear amplification reactions using a single primer, or polymerase chain reactions, employing two or more primers. Guidance for selecting the lengths and sequences of primers for particular applications is well known to those of ordinary skill in the art, as evidenced by the following references that are incorporated by reference: Dieffenbach, editor, *PCR Primer: A Laboratory Manual*, 2nd Edition (Cold Spring Harbor Press, New York, 2003).

[0084] “Quality score” means a measure of the probability that a base assignment at a particular sequence location is correct. A variety of methods are well known to those of ordinary skill for

calculating quality scores for particular circumstances, such as, for bases called as a result of different sequencing chemistries, detection systems, base-calling algorithms, and so on. Generally, quality score values are monotonically related to probabilities of correct base calling. For example, a quality score, or Q, of 10 may mean that there is a 90 percent chance that a base is called correctly, a Q of 20 may mean that there is a 99 percent chance that a base is called correctly, and so on. For some sequencing platforms, particularly those using sequencing-by-synthesis chemistries, average quality scores decrease as a function of sequence read length, so that quality scores at the beginning of a sequence read are higher than those at the end of a sequence read, such declines being due to phenomena such as incomplete extensions, carry forward extensions, loss of template, loss of polymerase, capping failures, deprotection failures, and the like.

[0085] “Repertoire”, or “immune repertoire”, means a set of distinct recombined nucleotide sequences that encode T cell receptors (TCRs) or B cell receptors (BCRs), or fragments thereof, respectively, in a population of lymphocytes of an individual, wherein the nucleotide sequences of the set have a one-to-one correspondence with distinct lymphocytes or their clonal subpopulations for substantially all of the lymphocytes of the population. . In one aspect, a population of lymphocytes from which a repertoire is determined is taken from one or more tissue samples, such as one or more blood samples. A member nucleotide sequence of a repertoire is referred to herein as a “clonotype.” In one aspect, clonotypes of a repertoire comprises any segment of nucleic acid common to a T cell or a B cell population which has undergone somatic recombination during the development of TCRs or BCRs, including normal or aberrant (e.g. associated with cancers) precursor molecules thereof, including, but not limited to, any of the following: an immunoglobulin heavy chain (IgH) or subsets thereof (e.g. an IgH variable region, CDR3 region, or the like), incomplete IgH molecules, an immunoglobulin light chain or subsets thereof (e.g. a variable region, CDR region, or the like), T cell receptor α chain or subsets thereof, T cell receptor β chain or subsets thereof (e.g. variable region, CDR3, V(D)J region, or the like), a CDR (including CDR1, CDR2 or CDR3, of either TCRs or BCRs, or combinations of such CDRs), V(D)J regions of either TCRs or BCRs, hypermutated regions of IgH variable regions, or the like. In one aspect, nucleic acid segments defining clonotypes of a repertoire are selected so that their diversity (i.e. the number of distinct nucleic acid sequences in the set) is large enough so that substantially every T cell or B cell or clone thereof in an individual carries a unique nucleic acid sequence of such repertoire. That is, in accordance with the invention, a practitioner may select for defining clonotypes a particular segment or region of recombined nucleic acids that encode TCRs or BCRs that do not reflect the full diversity of a

population of T cells or B cells; however, preferably, clonotypes are defined so that they do reflect the diversity of the population of T cells and/or B cells from which they are derived. That is, preferably each different clone of a sample has different clonotype. (Of course, in some applications, there will be multiple copies of one or more particular clonotypes within a profile, such as in the case of samples from leukemia or lymphoma patients). In other aspects of the invention, the population of lymphocytes corresponding to a repertoire may be circulating B cells, or may be circulating T cells, or may be subpopulations of either of the foregoing populations, including but not limited to, CD4+ T cells, or CD8+ T cells, or other subpopulations defined by cell surface markers, or the like. Such subpopulations may be acquired by taking samples from particular tissues, e.g. bone marrow, or lymph nodes, or the like, or by sorting or enriching cells from a sample (such as peripheral blood) based on one or more cell surface markers, size, morphology, or the like. In still other aspects, the population of lymphocytes corresponding to a repertoire may be derived from disease tissues, such as a tumor tissue, an infected tissue, or the like. In one embodiment, a repertoire comprising human TCR β chains or fragments thereof comprises a number of distinct nucleotide sequences in the range of from 0.1×10^6 to 1.8×10^6 , or in the range of from 0.5×10^6 to 1.5×10^6 , or in the range of from 0.8×10^6 to 1.2×10^6 . In another embodiment, a repertoire comprising human IgH chains or fragments thereof comprises a number of distinct nucleotide sequences in the range of from 0.1×10^6 to 1.8×10^6 , or in the range of from 0.5×10^6 to 1.5×10^6 , or in the range of from 0.8×10^6 to 1.2×10^6 . In a particular embodiment, a repertoire of the invention comprises a set of nucleotide sequences encoding substantially all segments of the V(D)J region of an IgH chain. In one aspect, “substantially all” as used herein means every segment having a relative abundance of .001 percent or higher; or in another aspect, “substantially all” as used herein means every segment having a relative abundance of .0001 percent or higher. In another particular embodiment, a repertoire of the invention comprises a set of nucleotide sequences that encodes substantially all segments of the V(D)J region of a TCR β chain. In another embodiment, a repertoire of the invention comprises a set of nucleotide sequences having lengths in the range of from 25-200 nucleotides and including segments of the V, D, and J regions of a TCR β chain. In another embodiment, a repertoire of the invention comprises a set of nucleotide sequences having lengths in the range of from 25-200 nucleotides and including segments of the V, D, and J regions of an IgH chain. In another embodiment, a repertoire of the invention comprises a number of distinct nucleotide sequences that is substantially equivalent to the number of lymphocytes expressing a distinct IgH chain. In another embodiment, a repertoire of the invention comprises a number of distinct nucleotide sequences that is substantially equivalent to

the number of lymphocytes expressing a distinct TCR β chain. In still another embodiment, “substantially equivalent” means that with ninety-nine percent probability a repertoire of nucleotide sequences will include a nucleotide sequence encoding an IgH or TCR β or portion thereof carried or expressed by every lymphocyte of a population of an individual at a frequency of .001 percent or greater. In still another embodiment, “substantially equivalent” means that with ninety-nine percent probability a repertoire of nucleotide sequences will include a nucleotide sequence encoding an IgH or TCR β or portion thereof carried or expressed by every lymphocyte present at a frequency of .0001 percent or greater. The sets of clonotypes described in the foregoing two sentences are sometimes referred to herein as representing the “full repertoire” of IgH and/or TCR β sequences. As mentioned above, when measuring or generating a clonotype profile (or repertoire profile), a sufficiently large sample of lymphocytes is obtained so that such profile provides a reasonably accurate representation of a repertoire for a particular application. In one aspect, samples comprising from 10^5 to 10^7 lymphocytes are employed, especially when obtained from peripheral blood samples of from 1-10 mL.

[0086] “Sequence read” means a sequence of nucleotides determined from a sequence or stream of data generated by a sequencing technique, which determination is made, for example, by means of base-calling software associated with the technique, e.g. base-calling software from a commercial provider of a DNA sequencing platform. A sequence read usually includes quality scores for each nucleotide in the sequence. Typically, sequence reads are made by extending a primer along a template nucleic acid, e.g. with a DNA polymerase or a DNA ligase. Data is generated by recording signals, such as optical, chemical (e.g. pH change), or electrical signals, associated with such extension. Such initial data is converted into a sequence read.

[0087] “Sequence tag” (or “tag”) or “barcode” means an oligonucleotide that is attached to a polynucleotide or template molecule and is used to identify and/or track the polynucleotide or template in a reaction or a series of reactions. A sequence tag may be attached to the 3’- or 5’- end of a polynucleotide or template or it may be inserted into the interior of such polynucleotide or template to form a linear conjugate, sometime referred to herein as a “tagged polynucleotide,” or “tagged template,” or “tag-polynucleotide conjugate,” “tag-molecule conjugate,” or the like. Sequence tags may vary widely in size and compositions; the following references, which are incorporated herein by reference, provide guidance for selecting sets of sequence tags appropriate for particular embodiments: Brenner, U.S. patent 5,635,400; Brenner and Macevicz, U.S. patent 7,537,897; Brenner et al, Proc. Natl. Acad. Sci., 97: 1665-1670 (2000); Church et al, European patent publication 0 303 459; Shoemaker et al, Nature Genetics, 14: 450-456 (1996); Morris et al, European patent publication 0799897A1; Wallace, U.S. patent 5,981,179; and the

like. Lengths and compositions of sequence tags can vary widely, and the selection of particular lengths and/or compositions depends on several factors including, without limitation, how tags are used to generate a readout, e.g. via a hybridization reaction or via an enzymatic reaction, such as sequencing; whether they are labeled, e.g. with a fluorescent dye or the like; the number of distinguishable oligonucleotide tags required to unambiguously identify a set of polynucleotides, and the like, and how different must tags of a set be in order to ensure reliable identification, e.g. freedom from cross hybridization or misidentification from sequencing errors. In one aspect, sequence tags can each have a length within a range of from 2 to 36 nucleotides, or from 4 to 30 nucleotides, or from 8 to 20 nucleotides, or from 6 to 10 nucleotides, respectively. In one aspect, sets of sequence tags are used wherein each sequence tag of a set has a unique nucleotide sequence that differs from that of every other tag of the same set by at least two bases; in another aspect, sets of sequence tags are used wherein the sequence of each tag of a set differs from that of every other tag of the same set by at least three bases.

[0088] “Sequence tree” means a tree data structure for representing nucleotide sequences. In one aspect, a tree data structure of the invention is a rooted directed tree comprising nodes and edges that do not include cycles, or cyclical pathways. Edges from nodes of tree data structures of the invention are usually ordered. Nodes and/or edges are structures that may contain, or be associated with, a value. Each node in a tree has zero or more child nodes, which by convention are shown below it in the tree. A node that has a child is called the child's parent node. A node has at most one parent. Nodes that do not have any children are called leaf nodes. The topmost node in a tree is called the root node. Being the topmost node, the root node will not have parents. It is the node at which operations on the tree commonly begin (although some algorithms begin with the leaf nodes and work up ending at the root). All other nodes can be reached from it by following edges or links.

What is claimed is:

1. A method for identifying lymphocytes that have infiltrated a solid tissue, the method comprising the steps of:

 sorting into one or more subsets a sample of lymphocytes from an accessible tissue of an individual;

 generating clonotype profiles for each of the one or more subsets of lymphocytes from the accessible tissue;

 generating at least one clonotype profile from at least one sample of the solid tissue; and

 detecting lymphocytes of each subset in the solid tissue from their respective clonotypes.

2. The method of claim 1 wherein at least one of said one or more subsets of lymphocytes comprise B cells, T cells, cytotoxic T cells, helper T cells, regulatory T cells, Th1 T cells, Th2 T cells, Th9 T cells, Th17 T cells, Tfh T cells, at least one subset of antigen-specific B cells and/or at least one subset of antigen-specific T cells.

3. The method of claim 2 wherein said step of detecting further includes determining numbers, levels, and/or ratios of lymphocytes of each of said one or more subsets.

4. The method of claim 3 wherein said solid tissue is a solid tumor.

5. The method of claim 4 wherein said one or more subsets of lymphocytes comprise regulatory T cells and cytotoxic T cells.

6. The method of claim 5 wherein said step of detecting includes determining a ratio of levels of regulation T cells to cytotoxic T cells in said solid tumor.

7. The method of claim 3 wherein said one or more subsets of lymphocytes comprise at least one subset of antigen-specific T cells.

8. The method of claim 1 wherein said accessible tissue is peripheral blood, bone marrow, lymph fluid, synovial fluid, or spinal cord fluid of said individual.
9. The method of claim 8 wherein said accessible tissue is peripheral blood of said individual.
10. The method of claim 1 wherein said step of detecting further includes enumerating said lymphocytes in each of said subsets by counting their respective clonotypes.
11. The method of claim 1 wherein said sample of said solid tissue comprises multiple tissue samples taken from different locations in said solid tissue.
12. The method of claim 1 wherein said one or more subsets of lymphocytes comprise helper T cells and cytotoxic T cells.
13. The method of claim 12 wherein said helper T cells are FoxP3⁺ regulatory T cells.
14. The method of claim 13 wherein said cytotoxic T cells are identified by a CD8⁺ marker and wherein said regulatory T cells are identified by markers CD4⁺, CD25^{+(high)}, and CD127^(low).
15. The method of claim 14 wherein said solid tissue is a solid tumor.
16. The method of claim 1 wherein said one or more subsets of lymphocytes are sorted into separate containers on the basis of one or more cell surface markers or intracellular markers.
17. The method of claim 16 wherein said step of sorting said one or more subsets of lymphocytes is implemented with a fluorescence-activated cell sorter (FACS).
18. The method of claim 1 wherein each of said clonotypes of said clonotype profiles comprises a recombined portion of a T-cell receptor gene or a recombined portion of an immunoglobulin gene.

19. The method of claim 18 wherein each of said clonotypes of said clonotype profiles comprises at least a portion of a V(D)J region of a nucleic acid encoding a T-cell receptor β chain or at least a portion of a V(D)J region of a nucleic acid encoding an IgH chain.
20. The method of claim 19 wherein each of said clonotype profiles comprises at least 1000 clonotypes of at least 30 nucleotides.
21. The method of claim 20 wherein each of said clonotype profiles comprises at least 10^4 clonotypes.
22. The method of claim 1 wherein said solid tissue is a solid tissue affected by an autoimmune disease.
23. The method of claim 22 wherein at least one of said one or more subsets of lymphocytes is selected from B cells, T cells, cytotoxic T cells, helper T cells, regulatory T cells, Th1 T cells, Th2 T cells, Th9 T cells, Th17 T cells, Tfh T cells, antigen-specific B cells and antigen-specific T cells, and wherein said step of detecting further includes determining numbers, levels, and/or ratios of lymphocytes of each of said one or more subsets.
24. The method of claim 23 wherein said one or more subsets of lymphocytes comprise regulatory T cells, Th1 T cells, Th2 T cells, Th9 T cells, Th17 T cells, Tfh T cells, and/or at least one subset of antigen-specific T cells.
25. The method of claim 23 wherein said solid tissue is selected from the group consisting of colon tissue, small intestine tissue, skin, connective tissue, subcutaneous tissue, lung tissue, kidney tissue.
26. The method of claim 1 wherein said solid tissue is a normal tissue, and wherein at least one of said one or more subsets of lymphocytes is selected from B cells, T cells, cytotoxic T cells, helper T cells, regulatory T cells, Th1 T cells, Th2 T cells, Th9 T cells, Th17 T cells, Tfh T cells, antigen-specific B cells and antigen-specific T cells, and wherein said step of detecting further includes determining numbers, levels, and/or ratios of lymphocytes of each of said one or more subsets.

27. A method for determining a prognosis from a state of lymphocyte infiltration into a solid tumor of a patient, the method comprising the steps of:

 sorting into one or more subsets a sample of lymphocytes from peripheral blood of the patient;

 generating clonotype profiles for each of the one or more subsets of lymphocytes from the peripheral blood;

 generating at least one clonotype profile from at least one sample of the solid tumor;
and

 determining numbers, levels, and/or ratios of lymphocytes of each of the one or more subsets.

28. The method of claim 27 wherein said one or more subsets include regulatory T cells and cytotoxic T cells.

29. The method of claim 28 wherein said step of determining includes indicating a greater than average rate of patient survival whenever a ratio of levels of said cytotoxic T cells to regulatory T cells is higher than that in a non-tumor tissue of said patient.