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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2010 (29.12.2010)

PCT

(10) International Publication Number
WO 2010/151416 A1

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

(21) International Application Number:
PCT/US2010/037477

(22) International Filing Date:
4 June 2010 (04.06.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/220,344 25 June 2009 (25.06.2009) US

(71) Applicant *(for all designated States except US)*: **FRED HUTCHINSON CANCER RESEARCH CENTER** [US/US]; 1100 Fairview Avenue North, Seattle, WA 98109 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ROBINS, Harlan, S.** [US/US]; 4418 Latona Avenue N.E., Seattle, WA 98105 (US). **WARREN, Edus, H.** [US/US]; 10770 N.E. Broomgerrie Road, Bainbridge Island, WA 98110 (US). **CARLSON, Christopher, Scott** [US/US]; 5125 107th Avenue N.E., Kirkland, WA 98033 (US).

(74) Agent: **KNUDSEN, Peter, J.**; Woodcock Washburn LLP, Cira Centre, 12th Floor, 2929 Arch Street, Philadelphia, PA 19104-2891 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

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

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

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

(54) Title: METHOD OF MEASURING ADAPTIVE IMMUNITY

(57) Abstract: A method of measuring immunocompetence is described. This method provides a means for assessing the effects of diseases or conditions that compromise the immune system and of therapies aimed to reconstitute it. This method is based on quantifying T-cell diversity by calculating the number of diverse T-cell receptor (TCR) beta chain variable regions from blood cells.

WO 2010/151416 A1

METHOD OF MEASURING ADAPTIVE IMMUNITY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 61/220,344, filed on June 25, 2009 and is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] What is described is a method to measure the adaptive immunity of a patient by analyzing the diversity of T cell receptor genes or antibody genes using large scale sequencing of nucleic acid extracted from adaptive immune system cells.

BACKGROUND

[0003] Immunocompetence is the ability of the body to produce a normal immune response (i.e., antibody production and/or cell-mediated immunity) following exposure to a pathogen, which might be a live organism (such as a bacterium or fungus), a virus, or specific antigenic components isolated from a pathogen and introduced in a vaccine. Immunocompetence is the opposite of immunodeficiency or immuno-incompetent or immunocompromised. Several examples would be a newborn that does not yet have a fully functioning immune system but may have maternally transmitted antibody (immunodeficient); a late stage AIDS patient with a failed or failing immune system (immuno-incompetent); a transplant recipient taking medication so their body will not reject the donated organ (immunocompromised); age-related attenuation of T cell function in the elderly; or individuals exposed to radiation or chemotherapeutic drugs. There may be cases of overlap but these terms are all indicators of a dysfunctional immune system. In reference to lymphocytes, immunocompetence means that a B cell or T cell is mature and can recognize antigens and allow a person to mount an immune response.

[0004] Immunocompetence depends on the ability of the adaptive immune system to mount an immune response specific for any potential foreign antigens, using the highly polymorphic receptors encoded by B cells (immunoglobulins, Igs) and T cells (T cell receptors, TCRs).

[0005] Igs expressed by B cells are proteins consisting of four polypeptide chains, two heavy chains (H chains) and two light chains (L chains), forming an H_2L_2 structure. Each pair of H and L chains contains a hypervariable domain, consisting of a V_L and a V_H region, and a constant domain. The H chains of Igs are of several types, μ, δ, γ, α, and β. The diversity of Igs within an individual is mainly determined by the hypervariable domain. The V domain of H chains is

created by the combinatorial joining of three types of germline gene segments, the V_H, D_H, and J_H segments. Hypervariable domain sequence diversity is further increased by independent addition and deletion of nucleotides at the V_H-D_H, D_H-J_H, and V_H-J_H junctions during the process of Ig gene rearrangement. In this respect, immunocompetence is reflected in the diversity of Igs.

[0006] TCRs expressed by αβ T cells are proteins consisting of two transmembrane polypeptide chains (α and β), expressed from the TCRA and TCRB genes, respectively. Similar TCR proteins are expressed in gamma-delta T cells, from the TCRD and TCRG loci. Each TCR peptide contains variable complementarity determining regions (CDRs), as well as framework regions (FRs) and a constant region. The sequence diversity of αβ T cells is largely determined by the amino acid sequence of the third complementarity-determining region (CDR3) loops of the α and β chain variable domains, which diversity is a result of recombination between variable (V_β), diversity (D_β), and joining (J_β) gene segments in the β chain locus, and between analogous V_α and J_α gene segments in the α chain locus, respectively. The existence of multiple such gene segments in the TCR α and β chain loci allows for a large number of distinct CDR3 sequences to be encoded. CDR3 sequence diversity is further increased by independent addition and deletion of nucleotides at the V_β-D_β, D_β-J_β, and V_α-J_α junctions during the process of TCR gene rearrangement. In this respect, immunocompetence is reflected in the diversity of TCRs.

[0007] There exists a long-felt need for methods of assessing or measuring the adaptive immune system of patients in a variety of settings, whether immunocompetence in the immunocompromised, or dysregulated adaptive immunity in autoimmune disease. A demand exists for methods of diagnosing a disease state or the effects of aging by assessing the immunocompetence of a patient. In the same way results of therapies that modify the immune system need to be monitored by assessing the immunocompetence of the patient while undergoing the treatment. Conversely, a demand exists for methods to monitor the adaptive immune system in the context of autoimmune disease flares and remissions, in order to monitor response to therapy, or the need to initiate prophylactic therapy pre-symptomatically.

SUMMARY

[0008] One aspect of the invention is composition comprising:

- a multiplicity of V-segment primers, wherein each primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and

- a multiplicity of J-segment primers, wherein each primer comprises a sequence that is complementary to a J segment;

wherein the V segment and J-segment primers permit amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of the TCR genes. One embodiment of the invention is the composition, wherein each V-segment primer comprises a sequence that is complementary to a single Vβ segment, and each J segment primer comprises a sequence that is complementary to a Jβ segment, and wherein V segment and J-segment primers permit amplification of a TCRβ CDR3 region. Another embodiment is the composition, wherein each V-segment primer comprises a sequence that is complementary to a single functional Vα segment, and each J segment primer comprises a sequence that is complementary to a Jα segment, and wherein V segment and J-segment primers permit amplification of a TCRα CDR3 region.

[0009] Another embodiment of the invention is the composition, wherein the V segment primers hybridize with a conserved segment, and have similar annealing strength. Another embodiment is wherein the V segment primer is anchored at position -43 in the Vβ segment relative to the recombination signal sequence (RSS). Another embodiment is wherein the multiplicity of V segment primers consist of at least 45 primers specific to 45 different Vβ genes. Another embodiment is wherein the V segment primers have sequences that are selected from the group consisting of SEQ ID NOS:1-45. Another embodiment is wherein the V segment primers have sequences that are selected from the group consisting of SEQ ID NOS:58-102. Another embodiment is wherein there is a V segment primer for each Vβ segment.

[0010] Another embodiment of the invention is the composition, wherein the J segment primers hybridize with a conserved framework region element of the Jβ segment, and have similar annealing strength. The composition of claim 2, wherein the multiplicity of J segment primers consist of at least thirteen primers specific to thirteen different Jβ genes. Another embodiment is The composition of claim 2, wherein the J segment primers have sequences that are selected from the group consisting of SEQ ID NOS:46-57. Another embodiment is wherein the J segment primers have sequences that are selected from the group consisting of SEQ ID NOS:102-113. Another embodiment is wherein there is a J segment primer for each Jβ segment. Another embodiment is wherein all J segment primers anneal to the same conserved motif.

[0011] Another embodiment of the invention is the composition, wherein the amplified DNA molecule starts from said conserved motif and amplifies adequate sequence to diagnostically identify the J segment and includes the CDR3 junction and extends into the V segment. Another

embodiment is wherein the amplified Jβ gene segments each have a unique four base tag at positions +11 through +14 downstream of the RSS site.

[0012] Another aspect of the invention is the composition further comprising a set of sequencing oligonucleotides, wherein the sequencing oligonucleotides hybridize to regions within the amplified DNA molecules. An embodiment is wherein the sequencing oligonucleotides hybridize adjacent to a four base tag within the amplified Jβ gene segments at positions +11 through +14 downstream of the RSS site. Another embodiment is wherein the sequencing oligonucleotides are selected from the group consisting of SEG ID NOS:58-70. Another embodiment is wherein the V-segment or J-segment are selected to contain a sequence error-correction by merger of closely related sequences. Another embodiment is the composition, further comprising a universal C segment primer for generating cDNA from mRNA.

[0013] Another aspect of the invention is a composition comprising:

- a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
- a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;

wherein the V segment and J segment primers permit amplification of the TCRG CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of antibody heavy chain genes.

[0014] Another aspect of the invention is a composition comprising:

- a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
- a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;

wherein the V segment and J segment primers permit amplification of antibody heavy chain (IGH) CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of antibody heavy chain genes.

[0015] Another aspect of the invention is a composition comprising:

- a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
- a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;

wherein the V segment and J segment primers permit amplification of antibody light chain (IGL) V_L region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of antibody light chain genes.

[0016] Another aspect of the invention is a method comprising:

- selecting a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
- selecting a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;
- combining the V segment and J segment primers with a sample of genomic DNA to permit amplification of a CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of the TCR genes.

[0017] One embodiment of the invention is the method wherein each V segment primer comprises a sequence that is complementary to a single functional Vβ segment, and each J segment primer comprises a sequence that is complementary to a Jβ segment; and wherein combining the V segment and J segment primers with a sample of genomic DNA permits amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) and produces a multiplicity of amplified DNA molecules. Another embodiment is wherein each V segment primer comprises a sequence that is complementary to a single functional Vα segment, and each J segment primer comprises a sequence that is complementary to a Jα segment; and wherein combining the V segment and J segment primers with a sample of genomic DNA permits amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) and produces a multiplicity of amplified DNA molecules.

[0018] Another embodiment of the invention is the method further comprising a step of sequencing the amplified DNA molecules. Another embodiment is wherein the sequencing step utilizes a set of sequencing oligonucleotides that hybridize to regions within the amplified DNA

molecules. Another embodiment is the method, further comprising a step of calculating the total diversity of TCRβ CDR3 sequences among the amplified DNA molecules. Another embodiment is wherein the method shows that the total diversity of a normal human subject is greater than $1*10^6$ sequences, greater than $2*10^6$ sequences, or greater than $3*10^6$ sequences.

[0019] Another aspect of the invention is a method of diagnosing immunodeficiency in a human patient, comprising measuring the diversity of TCR CDR3 sequences of the patient, and comparing the diversity of the subject to the diversity obtained from a normal subject. An embodiment of the invention is the method, wherein measuring the diversity of TCR sequences comprises the steps of:

- selecting a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
- selecting a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;
- combining the V segment and J segment primers with a sample of genomic DNA to permit amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules;
- sequencing the amplified DNA molecules;
- calculating the total diversity of TCR CDR3 sequences among the amplified DNA molecules.

[0020] An embodiment of the invention is the method, wherein comparing the diversity is determined by calculating using the following equation:

$$\Delta(t) = \sum_x E(n_x)_{measurement\,1+2} - \sum_x E(n_x)_{measurement\,2} = S\int_0^\infty e^{-\lambda}\left(1-e^{-\lambda t}\right)dG(\lambda)$$

wherein $G(\lambda)$ is the empirical distribution function of the parameters $\lambda_1, \ldots, \lambda_S$, n_x is the number of clonotypes sequenced exactly x times, and

$$E(n_x) = S\int_0^\infty \left(\frac{e^{-\lambda}\lambda^x}{x!}\right)dG(\lambda).$$

[0021] Another embodiment of the invention is the method, wherein the diversity of at least two samples of genomic DNA are compared. Another embodiment is wherein one sample of genomic DNA is from a patient and the other sample is from a normal subject. Another

embodiment is wherein one sample of genomic DNA is from a patient before a therapeutic treatment and the other sample is from the patient after treatment. Another embodiment is wherein the two samples of genomic DNA are from the same patient at different times during treatment. Another embodiment is wherein a disease is diagnosed based on the comparison of diversity among the samples of genomic DNA. Another embodiment is wherein the immunocompetence of a human patient is assessed by the comparison.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

[0022] The TCR and Ig genes can generate millions of distinct proteins via somatic mutation. Because of this diversity-generating mechanism, the hypervariable complementarity determining regions of these genes can encode sequences that can interact with millions of ligands, and these regions are linked to a constant region that can transmit a signal to the cell indicating binding of the protein's cognate ligand.

[0023] The adaptive immune system employs several strategies to generate a repertoire of T- and B-cell antigen receptors with sufficient diversity to recognize the universe of potential pathogens. In $\alpha\beta$ and $\gamma\delta$ T cells, which primarily recognize peptide antigens presented by MHC molecules, most of this receptor diversity is contained within the third complementarity-determining region (CDR3) of the T cell receptor (TCR) α and β chains (or γ and δ chains). Although it has been estimated that the adaptive immune system can generate up to 10^{18} distinct TCR $\alpha\beta$ pairs, direct experimental assessment of TCR CDR3 diversity has not been possible.

[0024] What is described herein is a novel method of measuring TCR CDR3 diversity that is based on single molecule DNA sequencing, and use this approach to sequence the CDR3 regions in millions of rearranged TCRβ genes isolated from peripheral blood T cells of two healthy adults.

[0025] The ability of the adaptive immune system to mount an immune response specific for any of the vast number of potential foreign antigens to which an individual might be exposed relies on the highly polymorphic receptors encoded by B cells (immunoglobulins) and T cells (T cell receptors; TCRs). The TCRs expressed by $\alpha\beta$ T cells, which primarily recognize peptide antigens presented by major histocompatibility complex (MHC) class I and II molecules, are heterodimeric proteins consisting of two transmembrane polypeptide chains (α and β), each containing one variable and one constant domain. The peptide specificity of $\alpha\beta$ T cells is in large part determined by the amino acid sequence encoded in the third complementarity-determining region (CDR3) loops of the α and β chain variable domains. The CDR3 regions of the β and α chains are formed by recombination between noncontiguous variable (V_β), diversity (D_β), and

joining (J_β) gene segments in the β chain locus, and between analogous V_α and J_α gene segments in the α chain locus, respectively. The existence of multiple such gene segments in the TCR α and β chain loci allows for a large number of distinct CDR3 sequences to be encoded. CDR3 sequence diversity is further increased by template-independent addition and deletion of nucleotides at the V_β-D_β, D_β-J_β, and V_α-J_α junctions during the process of TCR gene rearrangement.

[0026] Previous attempts to assess the diversity of receptors in the adult human $\alpha\beta$ T cell repertoire relied on examining rearranged TCR α and β chain genes expressed in small, well-defined subsets of the repertoire, followed by extrapolation of the diversity present in these subsets to the entire repertoire, to estimate approximately 10^6 unique TCRβ chain CDR3 sequences per individual, with 10-20% of these unique TCRβ CDR3 sequences expressed by cells in the antigen-experienced $CD45RO^+$ compartment. The accuracy and precision of this estimate is severely limited by the need to extrapolate the diversity observed in hundreds of sequences to the entire repertoire, and it is possible that the actual number of unique TCRβ chain CDR3 sequences in the $\alpha\beta$ T cell repertoire is significantly larger than 1×10^6.

[0027] Recent advances in high-throughput DNA sequencing technology have made possible significantly deeper sequencing than capillary-based technologies. A complex library of template molecules carrying universal PCR adapter sequences at each end is hybridized to a lawn of complementary oligonucleotides immobilized on a solid surface. Solid phase PCR is utilized to amplify the hybridized library, resulting in millions of template clusters on the surface, each comprising multiple (~1,000) identical copies of a single DNA molecule from the original library. A 30-54 bp interval in the molecules in each cluster is sequenced using reversible dye-termination chemistry, to permit simultaneous sequencing from genomic DNA of the rearranged TCRβ chain CDR3 regions carried in millions of T cells. This approach enables direct sequencing of a significant fraction of the uniquely rearranged TCRβ CDR3 regions in populations of $\alpha\beta$ T cells, which thereby permits estimation of the relative frequency of each CDR3 sequence in the population.

[0028] Accurate estimation of the diversity of TCRβ CDR3 sequences in the entire $\alpha\beta$ T cell repertoire from the diversity measured in a finite sample of T cells requires an estimate of the number of CDR3 sequences present in the repertoire that were not observed in the sample. TCRβ chain CDR3 diversity in the entire $\alpha\beta$ T cell repertoire were estimated using direct measurements of the number of unique TCRβ CDR3 sequences observed in blood samples containing millions of $\alpha\beta$ T cells. The results herein identify a lower bound for TCRβ CDR3 diversity in the $CD4^+$ and $CD8^+$ T cell compartments that is several fold higher than previous estimates. In addition,

the results herein demonstrate that there are at least 1.5 x 10^6 unique TCRβ CDR3 sequences in the CD45RO$^+$ compartment of antigen-experienced T-cells, a large proportion of which are present at low relative frequency. The existence of such a diverse population of TCRβ CDR3 sequences in antigen-experienced cells has not been previously demonstrated.

[0029] The diverse pool of TCRβ chains in each healthy individual is a sample from an estimated theoretical space of greater than 10^{11} possible sequences. However, the realized set of rearranged of TCRs is not evenly sampled from this theoretical space. Different Vβ's and Jβ's are found with over a thousand-fold frequency difference. Additionally, the insertion rates of nucleotides are strongly biased. This reduced space of realized TCRβ sequences leads to the possibility of shared β chains between people. With the sequence data generated by the methods described herein, the in vivo J usage, V usage, mono- and di- nucleotide biases, and position dependent amino acid usage can be computed. These biases significantly narrow the size of the sequence space from which TCRβ are selected, suggesting that different individuals share TCRβ chains with identical amino acid sequences. Results herein show that many thousands of such identical sequences are shared pairwise between individual human genomes.

[0030] The assay technology uses two pools of primers to provide for a highly multiplexed PCR reaction. The "forward" pool has a primer specific to each V segment in the gene (several primers targeting a highly conserved region are used, to simultaneously capture many V segments). The "reverse" pool primers anneal to a conserved sequence in the joining ("J") segment. The amplified segment pool includes adequate sequence to identify each J segment and also to allow for a J-segment-specific primer to anneal for resequencing. This enables direct observation of a large fraction of the somatic rearrangements present in an individual. This in turn enables rapid comparison of the TCR repertoire in individuals with an autoimmune disorder (or other target disease indication) against the TCR repertoire of controls.

[0031] The adaptive immune system can in theory generate an enormous diversity of T cell receptor CDR3 sequences – far more than are likely to be expressed in any one individual at any one time. Previous attempts to measure what fraction of this theoretical diversity is actually utilized in the adult $\alpha\beta$ T cell repertoire, however, have not permitted accurate assessment of the diversity. What is described herein is the development of a novel approach to this question that is based on single molecule DNA sequencing and an analytic computational approach to estimation of repertoire diversity using diversity measurements in finite samples. The analysis demonstrated that the number of unique TCRβ CDR3 sequences in the adult repertoire significantly exceeds previous estimates based on exhaustive capillary sequencing of small segments of the repertoire.

The TCRβ chain diversity in the $CD45RO^-$ population (enriched for naïve T cells) observed using the methods described herein is five-fold larger than previously reported. A major discovery is the number of unique TCRβ CDR3 sequences expressed in antigen-experienced $CD45RO^+$ T cells – the results herein show that this number is between 10 and 20 times larger than expected based on previous results of others. The frequency distribution of CDR3 sequences in $CD45RO^+$ cells suggests that the T cell repertoire contains a large number of clones with a small clone size.

[0032] The results herein show that the realized set of TCRβ chains are sampled non-uniformly from the huge potential space of sequences. In particular, the β chains sequences closer to germ line (few insertions and deletions at the V-D and D-J boundaries) appear to be created at a relatively high frequency. TCR sequences close to germ line are shared between different people because the germ line sequence for the V's, D's, and J's are shared, modulo a small number of polymorphisms, among the human population.

[0033] The T cell receptors expressed by mature αβ T cells are heterodimers whose two constituent chains are generated by independent rearrangement events of the TCR α and β chain variable loci. The α chain has less diversity than the β chain, so a higher fraction of α's are shared between individuals, and hundreds of exact TCR $\alpha\beta$ receptors are shared between any pair of individuals.

Cells

[0034] B cells and T cells can be obtained from a variety of tissue samples including marrow, thymus, lymph glands, peripheral tissues and blood, but peripheral blood is most easily accessed. Peripheral blood samples are obtained by phlebotomy from subjects. Peripheral blood mononuclear cells (PBMC) are isolated by techniques known to those of skill in the art, e.g., by Ficoll-Hypaque® density gradient separation. Preferably, whole PBMCs are used for analysis. The B and/or T lymphocytes, instead, may be flow sorted into multiple compartments for each subject: e.g. $CD8^+CD45RO^{+/-}$ and $CD4^+CD45RO^{+/-}$ using fluorescently labeled anti-human antibodies, e.g, CD4 FITC (clone M-T466, Miltenyi Biotec), CD8 PE (clone RPA-T8, BD Biosciences), CD45RO ECD (clone UCHL-1, Beckman Coulter), and CD45RO APC (clone UCHL-1, BD Biosciences). Staining of total PBMCs may be done with the appropriate combination of antibodies, followed by washing cells before analysis. Lymphocyte subsets can be isolated by FACS sorting, e.g., by a BD FACSAria™ cell-sorting system (BD Biosciences)

and by analyzing results with FlowJo software (Treestar Inc.), and also by conceptually similar methods involving specific antibodies immobilized to surfaces or beads.

Nucleic Acid Extraction

[0035] Total genomic DNA is extracted from cells, e.g., by using the QIAamp® DNA blood Mini Kit (QIAGEN®). The approximate mass of a single haploid genome is 3 pg. Preferably, at least 100,000 to 200,000 cells are used for analysis of diversity, i.e., about 0.6 to 1.2 μg DNA from diploid T cells. Using PBMCs as a source, the number of T cells can be estimated to be about 30% of total cells.

[0036] Alternatively, total nucleic acid can be isolated from cells, including both genomic DNA and mRNA. If diversity is to be measured from mRNA in the nucleic acid extract, the mRNA must be converted to cDNA prior to measurement. This can readily be done by methods of one of ordinary skill.

DNA Amplification

[0037] A multiplex PCR system is used to amplify rearranged TCR loci from genomic DNA, preferably from a CDR3 region, more preferably from a TCRα, TCRγ or TCRδ CDR3 region, most preferably from a TCRβ CDR3 region.

[0038] In general, a multiplex PCR system may use at least 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25, preferably 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, or 39, most preferably 40, 41, 42, 43, 44, or 45 forward primers, in which each forward primer is specific to a sequence corresponding to one or more TRB V region segments shown in SEQ ID NOS:114-248; and at least 3, 4, 5, 6, or 7, preferably 8, 9, 10, 11, 12 or 13 reverse primers, in which each reverse primer is specific to a sequence corresponding to one or more TRB J region segments shown in SEQ ID NOS:249-261. Most preferably, there is a J segment primer for every J segment.

[0039] Preferably, the primers are designed not to cross an intron/exon boundary. The forward primers must preferably anneal to the V segments in a region of relatively strong sequence conservation between V segments so as to maximize the conservation of sequence among these primers. Accordingly, this minimizes the potential for differential annealing properties of each primer, and so that the amplified region between V and J primers contains sufficient TCR V sequence information to identify the specific V gene segment used.

[0040] Preferably, the J segment primers hybridize with a conserved element of the J segment, and have similar annealing strength. Most preferably, all J segment primers anneal to the same conserved framework region motif. The forward and reverse primers are both preferably

modified at the 5' end with the universal forward primer sequence compatible with a DNA sequencer.

[0041] For example, a multiplex PCR system may use 45 forward primers (Table 1), each specific to a functional TCR Vβ segment, and thirteen reverse primers (Table 2), each specific to a TCR Jβ segment. Xn and Yn correspond to polynucleotides of lengths n and m, respectively, which would be specific to the single molecule sequencing technology being used to read out the assay.

Table 1: TCR-Vβ Forward primer sequences

TRBV gene segment(s)	SEQ ID NO:	Primer sequence*
TRBV2	1	XnTCAAATTTCACTCTGAAGATCCGGTCCACAA
TRBV3-1	2	XnGCTCACTTAAATCTTCACATCAATTCCCTGG
TRBV4-1	3	XnCTTAAACCTTCACCTACACGCCCTGC
TRBV(4-2, 4-3)	4	XnCTTATTCCTTCACCTACACACCCTGC
TRBV5-1	5	XnGCTCTGAGATGAATGTGAGCACCTTG
TRBV5-3	6	XnGCTCTGAGATGAATGTGAGTGCCTTG
TRBV(5-4, 5-5, 5-6, 5-7, 5-8)	7	XnGCTCTGAGCTGAATGTGAACGCCTTG
TRBV6-1	8	XnTCGCTCAGGCTGGAGTCGGCTG
TRBV(6-2, 6-3)	9	XnGCTGGGGTTGGAGTCGGCTG
TRBV6-4	10	XnCCCTCACGTTGGCGTCTGCTG
TRBV6-5	11	XnGCTCAGGCTGCTGTCGGCTG
TRBV6-6	12	XnCGCTCAGGCTGGAGTTGGCTG
TRBV6-7	13	XnCCCCTCAAGCTGGAGTCAGCTG
TRBV6-8	14	XnCACTCAGGCTGGTGTCGGCTG
TRBV6-9	15	XnCGCTCAGGCTGGAGTCAGCTG
TRBV7-1	16	XnCCACTCTGAAGTTCCAGCGCACAC
TRBV7-2	17	XnCACTCTGACGATCCAGCGCACAC
TRBV7-3	18	XnCTCTACTCTGAAGATCCAGCGCACAG
TRBV7-4	19	XnCCACTCTGAAGATCCAGCGCACAG
TRBV7-6	20	XnCACTCTGACGATCCAGCGCACAG
TRBV7-7	21	XnCCACTCTGACGATTCAGCGCACAG
TRBV7-8	22	XnCCACTCTGAAGATCCAGCGCACAC
TRBV7-9	23	XnCACCTTGGAGATCCAGCGCACAG
TRBV9	24	XnGCACTCTGAACTAAACCTGAGCTCTCTG
TRBV10-1	25	XnCCCCTCACTCTGGAGTCTGCTG
TRBV10-2	26	XnCCCCCTCACTCTGGAGTCAGCTA
TRBV10-3	27	XnCCTCCTCACTCTGGAGTCCGCTA
TRBV(11-1, 11-3)	28	XnCCACTCTCAAGATCCAGCCTGCAG
TRBV11-2	29	XnCTCCACTCTCAAGATCCAGCCTGCAA
TRBV(12-3, 12-4, 12-5)	30	XnCCACTCTGAAGATCCAGCCCTCAG
TRBV13	31	XnCATTCTGAACTGAACATGAGCTCCTTGG
TRBV14	32	XnCTACTCTGAAGGTGCAGCCTGCAG
TRBV15	33	XnGATAACTTCCAATCCAGGAGGCCGAACA
TRBV16	34	XnCTGTAGCCTTGAGATCCAGGCTACGA
TRBV17	35	XnCTTCCACGCTGAAGATCCATCCCG
TRBV18	36	XnGCATCCTGAGGATCCAGCAGGTAG
TRBV19	37	XnCCTCTCACTGTGACATCGGCCC
TRBV20-1	38	XnCTTGTCCACTCTGACAGTGACCAGTG
TRBV23-1	39	XnCAGCCTGGCAATCCTGTCCTCAG
TRBV24-1	40	XnCTCCCTGTCCCTAGAGTCTGCCAT
TRBV25-1	41	XnCCCTGACCCTGGAGTCTGCCA
TRBV27	42	XnCCCTGATCCTGGAGTCGCCCA
TRBV28	43	XnCTCCCTGATTCTGGAGTCCGCCA
TRBV29-1	44	XnCTAACATTCTCAACTCTGACTGTGAGCAACA
TRBV30	45	XnCGGCAGTTCATCCTGAGTTCTAAGAAGC

Table 2: TCR-Jβ Reverse Primer Sequences

TRBJ gene segment	SEQ ID NO:	Primer sequence*
TRBJ1-1	46	YmTTACCTACAACTGTGAGTCTGGTGCCTTGTCCAAA
TRBJ1-2	47	YmACCTACAACGGTTAACCTGGTCCCCGAACCGAA
TRBJ1-3	48	YmACCTACAACAGTGAGCCAACTTCCCTCTCCAAA
TRBJ1-4	49	YmCCAAGACAGAGAGCTGGGTTCCACTGCCAAA
TRBJ1-5	483	YmACCTAGGATGGAGAGTCGAGTCCCATCACCAAA
TRBJ1-6	50	YmCTGTCACAGTGAGCCTGGTCCCGTTCCCAAA
TRBJ2-1	51	YmCGGTGAGCCGTGTCCCTGGCCCGAA
TRBJ2-2	52	YmCCAGTACGGTCAGCCTAGAGCCTTCTCCAAA
TRBJ2-3	53	YmACTGTCAGCCGGGTGCCTGGGCCAAA
TRBJ2-4	54	YmAGAGCCGGGTCCCGGCGCCGAA
TRBJ2-5	55	YmGGAGCCGCGTGCCTGGCCCGAA
TRBJ2-6	56	YmGTCAGCCTGCTGCCGGCCCCGAA
TRBJ2-7	57	YmGTGAGCCTGGTGCCCGGCCCGAA

[0042] The 45 forward PCR primers of Table 1 are complementary to each of the 48 functional Variable segments, and the thirteen reverse PCR primers of Table 2 are complementary to each of the functional joining (J) gene segments from the TRB locus (TRBJ). The TRB V region segments are identified in the Sequence Listing at SEQ ID NOS:114-248 and the TRB J region segments are at SEQ ID NOS:249-261. The primers have been designed such that adequate information is present within the amplified sequence to identify both the V and J genes uniquely (>40 base pairs of sequence upstream of the V gene recombination signal sequence (RSS), and >30 base pairs downstream of the J gene RSS). Alternative primers may be selected by one of ordinary skill from the V and J regions of the genes of each TCR subunit.

[0043] The forward primers are modified at the 5' end with the universal forward primer sequence compatible with the DNA sequencer (Xn of Table 1). Similarly, all of the reverse primers are modified with a universal reverse primer sequence (Ym of Table 2). One example of such universal primers is shown in Tables 3 and 4, for the Illumina GAII single-end read sequencing system. The 45 TCR Vβ forward primers anneal to the Vβ segments in a region of relatively strong sequence conservation between Vβ segments so as to maximize the conservation of sequence among these primers.

Table 3: TCR-Vβ Forward primer sequences

TRBV gene segment(s)	SEQ ID NO:	Primer sequence*
TRBV2	58	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**TCAAATTTCACTCTGAAGATCCGGTCCACAA
TRBV3-1	59	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTCACTTAAATCTTCACATCAATTCCCTGG
TRBV4-1	60	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTTAAACCTTCACCTACACGCCCTGC
TRBV(4-2, 4-3)	61	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTTATTCCTTCACCTACACACCCTGC
TRBV5-1	62	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTCTGAGATGAATGTGAGCACCTTG
TRBV5-3	63	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTCTGAGATGAATGTGAGTGCCTTG
TRBV(5-4, 5-5, 5-6, 5-7, 5-8)	64	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTCTGAGCTGAATGTGAACGCCTTG
TRBV6-1	65	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**TCGCTCAGGCTGGAGTCGGCTG
TRBV(6-2, 6-3)	66	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTGGGGTTGGAGTCGGCTG
TRBV6-4	67	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCTCACGTTGGCGTCTGCTG
TRBV6-5	68	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCTCAGGCTGCTGTCGGCTG
TRBV6-6	69	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CGCTCAGGCTGGAGTTGGCTG
TRBV6-7	70	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCCTCAAGCTGGAGTCAGCTG
TRBV6-8	71	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CACTCAGGCTGGTGTCGGCTG
TRBV6-9	72	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CGCTCAGGCTGGAGTCAGCTG
TRBV7-1	73	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTGAAGTTCCAGCGCACAC
TRBV7-2	74	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CACTCTGACGATCCAGCGCACAC
TRBV7-3	75	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTCTACTCTGAAGATCCAGCGCACAG
TRBV7-4	76	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTGAAGATCCAGCGCACAG
TRBV7-6	77	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CACTCTGACGATCCAGCGCACAG
TRBV7-7	78	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTGACGATTCAGCGCACAG
TRBV7-8	79	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTGAAGATCCAGCGCACAC
TRBV7-9	80	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CACCTTGGAGATCCAGCGCACAG
TRBV9	81	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCACTCTGAACTAAACCTGAGCTCTCTG
TRBV10-1	82	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCCTCACTCTGGAGTCTGCTG
TRBV10-2	83	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCCCTCACTCTGGAGTCAGCTA
TRBV10-3	84	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCTCCTCACTCTGGAGTCCGCTA
TRBV(11-1, 11-3)	85	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTCAAGATCCAGCCTGCAG
TRBV11-2	86	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTCCACTCTCAAGATCCAGCCTGCAA
TRBV(12-3, 12-4, 12-5)	87	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCACTCTGAAGATCCAGCCCTCAG
TRBV13	88	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CATTCTGAACTGAACATGAGCTCCTTGG
TRBV14	89	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTACTCTGAAGGTGCAGCCTGCAG
TRBV15	90	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GATAACTTCCAATCCAGGAGGCCGAACA
TRBV16	91	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTGTAGCCTTGAGATCCAGGCTACGA
TRBV17	92	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTTCCACGCTGAAGATCCATCCCG
TRBV18	93	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**GCATCCTGAGGATCCAGCAGGTAG
TRBV19	94	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCTCTCACTGTGACATCGGCCC
TRBV20-1	95	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTTGTCCACTCTGACAGTGACCAGTG
TRBV23-1	96	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CAGCCTGGCAATCCTGTCCTCAG
TRBV24-1	97	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTCCCTGTCCCTAGAGTCTGCCAT
TRBV25-1	98	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCTGACCCTGGAGTCTGCCA
TRBV27	99	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CCCTGATCCTGGAGTCGCCCA
TRBV28	100	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTCCCTGATTCTGGAGTCCGCCA
TRBV29-1	101	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CTAACATTCTCAACTCTGACTGTGAGCAACA
TRBV30	102	**CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT**CGGCAGTTCATCCTGAGTTCTAAGAAGC

Table 4: TCR-*Jβ* Reverse Primer Sequences

TRBJ gene segment	SEQ ID NO:	Primer sequence*
TRBJ1-1	103	**AATGATACGGCGACCACCGAGATCT**TTACCTACAACTGTGAGTCTGGTGCCTTGTCCAAA
TRBJ1-2	468	**AATGATACGGCGACCACCGAGATCT**ACCTACAACGGTTAACCTGGTCCCCGAACCGAA
TRBJ1-3	104	**AATGATACGGCGACCACCGAGATCT**ACCTACAACAGTGAGCCAACTTCCCTCTCCAAA
TRBJ1-4	105	**AATGATACGGCGACCACCGAGATCT**CCAAGACAGAGAGCTGGGTTCCACTGCCAAA
TRBJ1-5	484	**AATGATACGGCGACCACCGAGATCT**ACCTAGGATGGAGAGTCGAGTCCCATCACCAAA
TRBJ1-6	106	**AATGATACGGCGACCACCGAGATCT**CTGTCACAGTGAGCCTGGTCCCGTTCCCAAA
TRBJ2-1	107	**AATGATACGGCGACCACCGAGATCT**CGGTGAGCCGTGTCCCTGGCCCGAA
TRBJ2-2	108	**AATGATACGGCGACCACCGAGATCT**CCAGTACGGTCAGCCTAGAGCCTTCTCCAAA
TRBJ2-3	109	**AATGATACGGCGACCACCGAGATCT**ACTGTCAGCCGGGTGCCTGGGCCAAA
TRBJ2-4	110	**AATGATACGGCGACCACCGAGATCT**AGAGCCGGGTCCCGGCGCCGAA
TRBJ2-5	111	**AATGATACGGCGACCACCGAGATCT**GGAGCCGCGTGCCTGGCCCGAA
TRBJ2-6	112	**AATGATACGGCGACCACCGAGATCT**GTCAGCCTGCTGCCGGCCCCGAA
TRBJ2-7	113	**AATGATACGGCGACCACCGAGATCT**GTGAGCCTGGTGCCCGGCCCGAA

* bold sequence indicates universal R oligonucleotide for the sequence analysis

[0044] The total PCR product for a rearranged TCRβ CDR3 region using this system is expected to be approximately 200 bp long. Genomic templates are PCR amplified using a pool of the 45 TCR Vβ F primers (the "VF pool") and a pool of the twleve TCR Jβ R primers (the "JR pool"). For example, 50 µl PCR reactions may be used with 1.0 µM VF pool (22 nM for each unique TCR Vβ F primer), 1.0 µM JR pool (77 nM for each unique TCRBJR primer), 1X QIAGEN Multiple PCR master mix (QIAGEN part number 206145), 10% Q-solution (QIAGEN), and 16 ng/ul gDNA.

[0045] The IGH primer set was designed to try to accommodate the potential for somatic hypermutation within the rearranged IGH genes, as is observed after initial stimulation of naïve B cells. Consequently all primers were designed to be slightly longer than normal, and to anchor the 3' ends of each primer into highly conserved sequences of three or more nucleotides that should be resistant to both functional and non-functional somatic mutations.

[0046] The IGHJ reverse primers were designed to anchor the 3' end of each PCR primer on a highly conserved GGGG sequence motif within the IGHJ segments. These sequences are shown in Table 5. Underlined sequence are ten base pairs in from RSS that may be deleted. These were excluded from barcode design. Bold sequence is the reverse complement of the IGH J reverse PCR primers. Italicized sequence is the barcode for J identity (eight barcodes reveal six genes, and two alleles within genes). Further sequence within underlined segment may reveal additional allelic identities.

Table 5

IgH J segment	SEQ ID NO:	Sequence
>IGHJ4*01/1-48	452	ACTACTTTGACTACTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCAG
>IGHJ4*03/1-48	453	GCTACTTTGACTACTGGGGCCAAGGGACCCTGGTCACCGTCTCCTCAG
>IGHJ4*02/1-48	454	ACTACTTTGACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
>IGHJ3*01/1-50	455	TGATGCTTTTGATGTCTGGGGCCAAGGGACAATGGTCACCGTCTCTTCAG
>IGHJ3*02/1-50	456	TGATGCTTTTGATATCTGGGGCCAAGGGACAATGGTCACCGTCTCTTCAG
>IGHJ6*01/1-63	457	ATTACTACTACTACTACGGTATGGACGTCTGGGGGCAAGGGACCACGGTCACCGTCTCCTCAG
>IGHJ6*02/1-62	458	ATTACTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCAG
>IGHJ6*04/1-63	459	ATTACTACTACTACTACGGTATGGACGTCTGGGGCAAAGGGACCACGGTCACCGTCTCCTCAG
>IGHJ6*03/1-62	460	ATTACTACTACTACTACTACATGGACGTCTGGGGCAAAGGGACCACGGTCACCGTCTCCTCAG
>IGHJ2*01/1-53	461	CTACTGGTACTTCGATCTCTGGGGCCGTGGCACCCTGGTCACTGTCTCCTCAG
>IGHJ5*01/1-51	462	ACAACTGGTTCGACTCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCAG
>IGHJ5*02/1-51	463	ACAACTGGTTCGACCCCTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
>IGHJ1*01/1-52	464	GCTGAATACTTCCAGCACTGGGGCCAGGGCACCCTGGTCACCGTCTCCTCAG
>IGHJ2P*01/1-61	465	CTACAAGTGCTTGGAGCACTGGGGCAGGGCAGCCCGGACACCGTCTCCCTGGGAACGTCAG
>IGHJ1P*01/1-54	466	AAAGGTGCTGGGGGTCCCCTGAACCCGACCCGCCCTGAGACCGCAGCCACATCA
>IGHJ3P*01/1-52	467	CTTGCGGTTGGACTTCCCAGCCGACAGTGGTGGTCTGGCTTCTGAGGGGTCA

Sequences of the IGHJ reverse PCR primers are shown in Table 6.

Table 6

IgH J segment	SEQ ID NO:	sequence
>IGHJ4_1	421	TGAGGAGACGGTGACCAGGGTTCCTTGGCCC
>IGHJ4_3	422	TGAGGAGACGGTGACCAGGGTCCCTTGGCCC
>IGHJ4_2	423	TGAGGAGACGGTGACCAGGGTTCCCTGGCCC
>IGHJ3_12	424	CTGAAGAGACGGTGACCATTGTCCCTTGGCCC
>IGHJ6_1	425	CTGAGGAGACGGTGACCGTGGTCCCTTGCCCC
>IGHJ6_2	426	TGAGGAGACGGTGACCGTGGTCCCTTGGCCC
>IGHJ6_34	427	CTGAGGAGACGGTGACCGTGGTCCCTTTGCCC
>IGHJ2_1	428	CTGAGGAGACAGTGACCAGGGTGCCACGGCCC
>IGHJ5_1	429	CTGAGGAGACGGTGACCAGGGTTCCTTGGCCC
>IGHJ5_2	430	CTGAGGAGACGGTGACCAGGGTTCCCTGGCCC
>IGHJ1_1	431	CTGAGGAGACGGTGACCAGGGTGCCCTGGCCC

[0047] V primers were designed in a conserved in region of FR2 between the two conserved tryptophan (W) codons.

[0048] The primer sequences are anchored at the 3' end on a tryptophan codon for all IGHV families that conserve this codon. his allows for the last three nucleotides (tryptophan's TGG) to anchor on sequence that is expected to be resistant to somatic hypermutation, providing a 3'

anchor of five out of six nucleotides for each primer. The upstream sequence is extended further than normal, and includes degenerate nucleotides to allow for mismatches induced by hypermutation (or between closely relate IGH V families) without dramatically changing the annealing characteristics of the primer, as shown in Table 7. The sequences of the V gene segments are SEQ ID NOS:262-420.

Table 7

IgH V segment	SEQ ID NO:	sequence
>IGHV1	443	TGGGTGCACCAGGTCCANGNACAAGGGCTTGAGTGG
>IGHV2	444	TGGGTGCGACAGGCTCGNGNACAACGCCTTGAGTGG
>IGHV3	445	TGGGTGCGCCAGATGCCNGNGAAAGGCCTGGAGTGG
>IGHV4	446	TGGGTCCGCCAGSCYCCNGNGAAGGGGCTGGAGTGG
>IGHV5	447	TGGGTCCGCCAGGCTCCNGNAAAGGGGCTGGAGTGG
>IGHV6	448	TGGGTCTGCCAGGCTCCNGNGAAGGGGCAGGAGTGG
>IGH7_3.25p	449	TGTGTCCGCCAGGCTCCAGGGAATGGGCTGGAGTTGG
>IGH8_3.54p	450	TCAGATTCCCAAGCTCCAGGGAAGGGGCTGGAGTGAG
>IGH9_3.63p	451	TGGGTCAATGAGACTCTAGGGAAGGGGCTGGAGGGAG

[0049] Thermal cycling conditions may follow methods of those skilled in the art. For example, using a PCR Express thermal cycler (Hybaid, Ashford, UK), the following cycling conditions may be used: 1 cycle at 95°C for 15 minutes, 25 to 40 cycles at 94°C for 30 seconds, 59°C for 30 seconds and 72°C for 1 minute, followed by one cycle at 72°C for 10 minutes.

Sequencing

[0050] Sequencing is achieved using a set of sequencing oligonucleotides that hybridize to a defined region within the amplified DNA molecules.

[0051] Preferably, the amplified J gene segments each have a unique four base tag at positions +11 through +14 downstream from the RSS site. Accordingly, the sequencing oligonucleotides hybridize adjacent to a four base tag within the amplified Jβ gene segments at positions +11 through +14 downstream of the RSS site.

[0052] For example, sequencing oligonucleotides for TCRB may be designed to anneal to a consensus nucleotide motif observed just downstream of this "tag", so that the first four bases of a sequence read will uniquely identify the J segment (Table 8).

Table 8: Sequencing oligonucleotides

Sequencing oligonucleotide	SEQ ID NO:	Oligonucleotide sequence
Jseq 1-1	470	ACAACTGTGAGTCTGGTGCCTTGTCCAAAGAAA
Jseq 1-2	471	ACAACGGTTAACCTGGTCCCCGAACCGAAGGTG
Jseq 1-3	472	ACAACAGTGAGCCAACTTCCCTCTCCAAAATAT
Jseq 1-4	473	AAGACAGAGAGCTGGGTTCCACTGCCAAAAAAC
Jseq 1-5	474	AGGATGGAGAGTCGAGTCCCATCACCAAAATGC
Jseq 1-6	475	GTCACAGTGAGCCTGGTCCCGTTCCCAAAGTGG
Jseq 2-1	476	AGCACGGTGAGCCGTGTCCCTGGCCCGAAGAAC
Jseq 2-2	477	AGTACGGTCAGCCTAGAGCCTTCTCCAAAAAAC
Jseq 2-3	478	AGCACTGTCAGCCGGGTGCCTGGGCCAAAATAC
Jseq 2-4	479	AGCACTGAGAGCCGGGTCCCGGCGCCGAAGTAC
Jseq 2-5	480	AGCACCAGGAGCCGCGTGCCTGGCCCGAAGTAC
Jseq 2-6	481	AGCACGGTCAGCCTGCTGCCGGCCCCGAAAGTC
Jseq 2-7	482	GTGACCGTGAGCCTGGTGCCCGGCCCGAAGTAC

[0053] The information used to assign the J and V segment of a sequence read is entirely contained within the amplified sequence, and does not rely upon the identity of the PCR primers. These sequencing oligonucleotides were selected such that promiscuous priming of a sequencing reaction for one J segment by an oligonucleotide specific to another J segment would generate sequence data starting at exactly the same nucleotide as sequence data from the correct sequencing oligonucleotide. In this way, promiscuous annealing of the sequencing oligonucleotides did not impact the quality of the sequence data generated.

[0054] The average length of the CDR3 region, defined as the nucleotides between the second conserved cysteine of the V segment and the conserved phenylalanine of the J segment, is 35+/-3, so sequences starting from the Jβ segment tag will nearly always capture the complete V-D-J junction in a 50 base pair read.

[0055] TCR βJ gene segments are roughly 50 base pair in length. PCR primers that anneal and extend to mismatched sequences are referred to as promiscuous primers. The TCR Jβ Reverse PCR primers were designed to minimize overlap with the sequencing oligonucleotides to minimize promiscuous priming in the context of multiplex PCR. The 13 TCR Jβ reverse primers are anchored at the 3' end on the consensus splice site motif, with minimal overlap of the sequencing primers. The TCR Jβ primers provide consistent annealing temperature using the sequencer program under default parameters.

[0056] For the sequencing reaction, the IGHJ sequencing primers extend three nucleotides across the conserved CAG sequences as shown in Table 9.

Table 9

IgH J segment	SEQ ID NO:	sequence
>IGHJSEQ4_1	432	TGAGGAGACGGTGACCAGGGTTCCTTGGCCCCAG
>IGHJSEQ4_3	433	TGAGGAGACGGTGACCAGGGTCCCTTGGCCCCAG
>IGHJSEQ4_2	434	TGAGGAGACGGTGACCAGGGTTCCCTGGCCCCAG
>IGHJSEQ3_12	435	CTGAAGAGACGGTGACCATTGTCCCTTGGCCCCAG
>IGHJSEQ6_1	436	CTGAGGAGACGGTGACCGTGGTCCCTTGCCCCCAG
>IGHJSEQ6_2	437	TGAGGAGACGGTGACCGTGGTCCCTTGGCCCCAG
>IGHJSEQ6_34	438	CTGAGGAGACGGTGACCGTGGTCCCTTTGCCCCAG
>IGHJSEQ2_1	439	CTGAGGAGACAGTGACCAGGGTGCCACGGCCCCAG
>IGHJSEQ5_1	440	CTGAGGAGACGGTGACCAGGGTTCCTTGGCCCCAG
>IGHJSEQ5_2	441	CTGAGGAGACGGTGACCAGGGTTCCCTGGCCCCAG
>IGHJSEQ1_1	442	CTGAGGAGACGGTGACCAGGGTGCCCTGGCCCCAG

Processing sequence data

[0057] For rapid analysis of sequencing results, an algorithm can be developed by one of ordinary skill. A preferred method is as follows.

[0058] The use of a PCR step to amplify the TCRβ CDR3 regions prior to sequencing could potentially introduce a systematic bias in the inferred relative abundance of the sequences, due to differences in the efficiency of PCR amplification of CDR3 regions utilizing different Vβ and Jβ gene segments. Each cycle of PCR amplification potentially introduces a bias of average magnitude $1.5^{1/15} = 1.027$. Thus, the 25 cycles of PCR introduces a total bias of average magnitude $1.027^{25} = 1.95$ in the inferred relative abundance of distinct CDR3 region sequences.

[0059] Sequenced reads were filtered for those including CDR3 sequences. Sequencer data processing involves a series of steps to remove errors in the primary sequence of each read, and to compress the data. A complexity filter removes approximately 20% of the sequences that are misreads from the sequencer. Then, sequences were required to have a minimum of a six base match to both one of the thirteen TCRB J-regions and one of 54 V-regions. Applying the filter to the control lane containing phage sequence, on average only one sequence in 7-8 million passed these steps. Finally, a nearest neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences, in order to remove both PCR error and sequencing error.

[0060] Analyzing the data, the ratio of sequences in the PCR product must be derived working backward from the sequence data before estimating the true distribution of clonotypes in the blood. For each sequence observed a given number of times in the data herein, the probability that that sequence was sampled from a particular size PCR pool is estimated. Because the CDR3 regions sequenced are sampled randomly from a massive pool of PCR products, the number of

observations for each sequence are drawn from Poisson distributions. The Poisson parameters are quantized according to the number of T cell genomes that provided the template for PCR. A simple Poisson mixture model both estimates these parameters and places a pairwise probability for each sequence being drawn from each distribution. This is an expectation maximization method which reconstructs the abundances of each sequence that was drawn from the blood.

[0061] To estimate diversity, the "unseen species" formula is employed. To apply this formula, unique adaptive immune receptors (e.g. TCRB) clonotypes takes the place of species. The mathematical solution provides that for a total number of TCRβ "species" or clonotypes, S, a sequencing experiment observes x_s copies of sequence s. For all of the unobserved clonotypes, x_s equals 0, and each TCR clonotype is "captured" in a blood draw according to a Poisson process with parameter λ_s. The number of T cell genomes sequenced in the first measurement 1, and in the second measurement. Since there are a large number of unique sequences, an integral will represent the sum. If $G(\lambda)$ is the empirical distribution function of the parameters $\lambda_1, \ldots, \lambda_S$, and n_x is the number of clonotypes sequenced exactly x times, then the total number of clonotypes, i.e., the measurement of diversity E, is given by the following formula:

$$E(n_x) = S\int_0^\infty \left(\frac{e^{-\lambda}\lambda^x}{x!}\right) dG(\lambda).$$

[0062] For a given experiment, where T cells are sampled from some arbitrary source (e.g. a blood draw), the formula is used to estimate the total diversity of species in the entire source. The idea is that the sampled number of clonotypes at each size contains sufficient information to estimate the underlying distribution of clonotypes in the whole source. To derive the formula, the number of new species expected if the exact measurement was repeated was estimated. The limit of the formula as if repeating the measurements an infinite number of times. The result is the expect number of species in the total underlying source population. The value for Δ(t), the number of *new* clonotypes observed in a second measurement, should be determined, preferably using the following equation:

$$\Delta(t) = \sum_x E(n_x)_{msmt1+msmt2} - \sum_x E(n_x)_{msmt1} = S\int_0^\infty e^{-\lambda}\left(1-e^{-\lambda t}\right) dG(\lambda)$$

in which *msmt1* and *msmt2* are the number of clonotypes from measurement 1 and 2, respectively. Taylor expansion of $1-e^{-\lambda t}$ gives $\Delta(t) = E(x_1)t - E(x_2)t^2 + E(x_3)t^3 - \ldots$, which can be approximated by replacing the expectations $E(n_x)$ with the observed numbers in the first

measurement. Using in the numbers observed in the first measurement, this formula predicts that $1.6*10^5$ new unique sequences should be observed in the second measurement. The actual value of the second measurement was $1.8*10^5$ new TCRβ sequences, which implies that the prediction provided a valid lower bound on total diversity. An Euler's transformation was used to regularize $\Delta(t)$ to produce a lower bound for $\Delta(\infty)$.

Using a measurement of diversity to diagnose disease

[0063] The measurement of diversity can be used to diagnose disease or the effects of a treatment, as follows. T cell and/or B cell receptor repertoires can be measured at various time points, e.g., after hematopoietic stem cell transplant (HSCT) treatment for leukemia. Both the change in diversity and the overall diversity of TCRB repertoire can be utilized to measure immunocompetence. A standard for the expected rate of immune reconstitution after transplant can be utilized. The rate of change in diversity between any two time points may be used to actively modify treatment. The overall diversity at a fixed time point is also an important measure, as this standard can be used to compare between different patients. In particular, the overall diversity is the measure that should correlate with the clinical definition of immune reconstitution. This information may be used to modify prophylactic drug regiments of antibiotics, antivirals, and antifungals, e.g., after HSCT.

[0064] The assessment of immune reconstitution after allogeneic hematopoietic cell transplantation can be determined by measuring changes in diversity. These techniques will also enhance the analysis of how lymphocyte diversity declines with age, as measured by analysis of T cell responses to vaccination. Further, the methods of the invention provide a means to evaluate investigational therapeutic agents (e.g., Interleukin-7 (IL-7)) that have a direct effect on the generation, growth, and development of αβ T cells. Moreover, application of these techniques to the study of thymic T cell populations will provide insight into the processes of both T cell receptor gene rearrangement as well as positive and negative selection of thymocytes.

[0065] A newborn that does not yet have a fully functioning immune system but may have maternally transmitted antibody is immunodeficient. A newborn is susceptible to a number of diseases until its immune system autonomously develops, and our measurement of the adaptive immune system may will likely prove useful with newborn patients.

[0066] Lymphocyte diversity can be assessed in other states of congenital or acquired immunodeficiency. An AIDS patient with a failed or failing immune system can be monitored to

determine the stage of disease, and to measure a patient's response to therapies aimed to reconstitute immunocompetence.

[0067] Another application of the methods of the invention is to provide diagnostic measures for solid organ transplant recipients taking medication so their body will not reject the donated organ. Generally, these patients are under immunosuppressive therapies. Monitoring the immunocompetence of the host will assist before and after transplantation.

[0068] Individuals exposed to radiation or chemotherapeutic drugs are subject to bone marrow transplantations or otherwise require replenishment of T cell populations, along with associated immunocompetence. The methods of the invention provide a means for qualitatively and quantitatively assessing the bone marrow graft, or reconstitution of lymphocytes in the course of these treatments.

[0069] One manner of determining diversity is by comparing at least two samples of genomic DNA, preferably in which one sample of genomic DNA is from a patient and the other sample is from a normal subject, or alternatively, in which one sample of genomic DNA is from a patient before a therapeutic treatment and the other sample is from the patient after treatment, or in which the two samples of genomic DNA are from the same patient at different times during treatment. Another manner of diagnosis may be based on the comparison of diversity among the samples of genomic DNA, e.g., in which the immunocompetence of a human patient is assessed by the comparison.

Biomarkers

[0070] Shared TCR sequences between individuals represent a new class of potential biomarkers for a variety of diseases, including cancers, autoimmune diseases, and infectious diseases. These are the public T cells that have been reported for multiple human diseases. TCRs are useful as biomarkers because T cells are a result of clonal expansion, by which the immune system amplifies these biomarkers through rapid cell division. Following amplification, the TCRs are readily detected even if the target is small (e.g. an early stage tumor). TCRs are also useful as biomarkers because in many cases the T cells might additionally contribute to the disease causally and, therefore could constitute a drug target. T cells self interactions are thought to play a major role in several diseases associated with autoimmunity, e.g., multiple sclerosis, Type I diabetes, and rheumatoid arthritis.

EXAMPLES

[0071] **Example 1**: Sample acquisition, PBMC isolation, FACS sorting and genomic DNA extraction

[0072] Peripheral blood samples from two healthy male donors aged 35 and 37 were obtained with written informed consent using forms approved by the Institutional Review Board of the Fred Hutchinson Cancer Research Center (FHCRC). Peripheral blood mononuclear cells (PBMC) were isolated by Ficoll-Hypaque® density gradient separation. The T-lymphocytes were flow sorted into four compartments for each subject: $CD8^+CD45RO^{+/-}$ and $CD4^+CD45RO^{+/-}$. For the characterization of lymphocytes the following conjugated anti-human antibodies were used: CD4 FITC (clone M-T466, Miltenyi Biotec), CD8 PE (clone RPA-T8, BD Biosciences), CD45RO ECD (clone UCHL-1, Beckman Coulter), and CD45RO APC (clone UCHL-1, BD Biosciences). Staining of total PBMCs was done with the appropriate combination of antibodies for 20 minutes at 4°C, and stained cells were washed once before analysis. Lymphocyte subsets were isolated by FACS sorting in the BD FACSAria™ cell-sorting system (BD Biosciences). Data were analyzed with FlowJo software (Treestar Inc.).

[0073] Total genomic DNA was extracted from sorted cells using the QIAamp® DNA blood Mini Kit (QIAGEN®). The approximate mass of a single haploid genome is 3 pg. In order to sample millions of rearranged TCRB in each T cell compartment, 6 to 27 micrograms of template DNA were obtained from each compartment (see Table 10).

Table 10

	CD8+/CD 45RO-	CD8+/CD45RO+	CD4+/CD45RO-	CD4+/CD45RO+	Donor
cells ($x10^6$)	9.9	6.3	6.3	10	
DNA (μg)	27	13	19	25	
PCR cycles	25	25	30	30	2
clusters (K/tile)	29.3	27	102.3*	118.3*	
VJ sequences ($x10^6$)	3.0	2.0	4.4	4.2	
Cells	4.9	4.8	3.3	9	
DNA	12	13	6.6	19	
PCR cycles	30	30	30	30	1
Clusters	116.3	121	119.5	124.6	
VJ sequences	3.2	3.7	4.0	3.8	
Cells	NA	NA	NA	0.03	
DNA	NA	NA	NA	0.015	
PCR cycles	NA	NA	NA	25 + 15	PCR Bias assessment
clusters	NA	NA	NA	1.4 / 23.8	
VJ sequences	NA	NA	NA	1.6	

[0074] **Example 2**: Virtual T cell receptor β chain spectratyping

[0075] Virtual TCR β chain spectratyping was performed as follows. Complementary DNA was synthesized from RNA extracted from sorted T cell populations and used as template for multiplex PCR amplification of the rearranged TCR β chain CDR3 region. Each multiplex reaction contained a 6-FAM-labeled antisense primer specific for the TCR β chain constant region, and two to five TCR β chain variable (TRBV) gene-specific sense primers. All 23 functional Vβ families were studied. PCR reactions were carried out on a Hybaid PCR Express thermal cycler (Hybaid, Ashford, UK) under the following cycling conditions: 1 cycle at 95°C for 6 minutes, 40 cycles at 94°C for 30 seconds, 58°C for 30 seconds, and 72°C for 40 seconds, followed by 1 cycle at 72°C for 10 minutes. Each reaction contained cDNA template, 500 μM dNTPs, 2mM $MgCl_2$ and 1 unit of AmpliTaq Gold DNA polymerase (Perkin Elmer) in AmpliTaq Gold buffer, in a final volume of 20 μl. After completion, an aliquot of the PCR

product was diluted 1:50 and analyzed using a DNA analyzer. The output of the DNA analyzer was converted to a distribution of fluorescence intensity vs. length by comparison with the fluorescence intensity trace of a reference sample containing known size standards.

[0076] **Example 3**: Multiplex PCR amplification of TCRβ CDR3 regions

[0077] The CDR3 junction region was defined operationally, as follows. The junction begins with the second conserved cysteine of the V-region and ends with the conserved phenylalanine of the J-region. Taking the reverse complements of the observed sequences and translating the flanking regions, the amino acids defining the junction boundaries were identified. The number of nucleotides between these boundaries determines the length and therefore the frame of the CDR3 region.In order to generate the template library for sequencing, a multiplex PCR system was selected to amplify rearranged TCR*β* loci from genomic DNA. The multiplex PCR system uses 45 forward primers (Table 3), each specific to a functional TCR V*β* segment, and thirteen reverse primers (Table 4), each specific to a TCR J*β* segment. The primers were selected to provide that adequate information is present within the amplified sequence to identify both the V and J genes uniquely (>40 base pairs of sequence upstream of the V gene recombination signal sequence (RSS), and >30 base pairs downstream of the J gene RSS).

[0078] The forward primers are modified at the 5' end with the universal forward primer sequence compatible with the Illumina GA2 cluster station solid-phase PCR. Similarly, all of the reverse primers are modified with the GA2 universal reverse primer sequence. The 3' end of each forward primer is anchored at position -43 in the Vβ segment, relative to the recombination signal sequence (RSS), thereby providing a unique Vβ tag sequence within the amplified region. The thirteen reverse primers specific to each Jβ segment are anchored in the 3' intron, with the 3' end of each primer crossing the intron/exon junction. Thirteen sequencing primers complementary to the Jβ segments were designed that are complementary to the amplified portion of the *J*β segment, such that the first few bases of sequence generated will capture the unique Jβ tag sequence.

[0079] On average J deletions were 4 bp +/- 2.5 bp, which implies that J deletions greater than 10 nucleotides occur in less than 1% of sequences. The thirteen different TCR J*β* gene segments each had a unique four base tag at positions +11 through +14 downstream of the RSS site. Thus, sequencing oligonucleotides were designed to anneal to a consensus nucleotide motif observed just downstream of this "tag", so that the first four bases of a sequence read will uniquely identify the J segment (Table 5).

[0080] The information used to assign the J and V segment of a sequence read is entirely contained within the amplified sequence, and does not rely upon the identity of the PCR primers. These sequencing oligonucleotides were selected such that promiscuous priming of a sequencing reaction for one J segment by an oligonucleotide specific to another J segment would generate sequence data starting at exactly the same nucleotide as sequence data from the correct sequencing oligonucleotide. In this way, promiscuous annealing of the sequencing oligonucleotides did not impact the quality of the sequence data generated.

[0081] The average length of the CDR3 region, defined following convention as the nucleotides between the second conserved cysteine of the V segment and the conserved phenylalanine of the J segment, is 35+/-3, so sequences starting from the Jβ segment tag will nearly always capture the complete VNDNJ junction in a 50 bp read.

[0082] TCR βJ gene segments are roughly 50 bp in length. PCR primers that anneal and extend to mismatched sequences are referred to as promiscuous primers. Because of the risk of promiscuous priming in the context of multiplex PCR, especially in the context of a gene family, the TCR Jβ Reverse PCR primers were designed to minimize overlap with the sequencing oligonucleotides. Thus, the 13 TCR Jβ reverse primers are anchored at the 3' end on the consensus splice site motif, with minimal overlap of the sequencing primers. The TCR Jβ primers were designed for a consistent annealing temperature (58 degrees in 50 mM salt) using the OligoCalc program under default parameters (http://www.basic.northwestern.edu/biotools/oligocalc.html).

[0083] The 45 TCR Vβ forward primers were designed to anneal to the Vβ segments in a region of relatively strong sequence conservation between Vβ segments, for two express purposes. First, maximizing the conservation of sequence among these primers minimizes the potential for differential annealing properties of each primer. Second, the primers were chosen such that the amplified region between V and J primers will contain sufficient TCR V*β* sequence information to identify the specific Vβ gene segment used. This obviates the risk of erroneous TCR Vβ gene segment assignment, in the event of promiscuous priming by the TCR Vβ primers. TCR Vβ forward primers were designed for all known non-pseudogenes in the TCR*β* locus.

[0084] The total PCR product for a successfully rearranged TCRβ CDR3 region using this system is expected to be approximately 200 bp long. Genomic templates were PCR amplified using an equimolar pool of the 45 TCR Vβ F primers (the "VF pool") and an equimolar pool of the thirteen TCR Jβ R primers (the "JR pool"). 50 µl PCR reactions were set up at 1.0 µM VF pool (22 nM for each unique TCR Vβ F primer), 1.0 µM JR pool (77 nM for each unique

TCRBJR primer), 1X QIAGEN Multiple PCR master mix (QIAGEN part number 206145), 10% Q-solution (QIAGEN), and 16 ng/ul gDNA. The following thermal cycling conditions were used in a PCR Express thermal cycler (Hybaid, Ashford, UK) under the following cycling conditions: 1 cycle at 95°C for 15 minutes, 25 to 40 cycles at 94°C for 30 seconds, 59°C for 30 seconds and 72°C for 1 minute, followed by one cycle at 72°C for 10 minutes. 12-20 wells of PCR were performed for each library, in order to sample hundreds of thousands to millions of rearranged TCRβ CDR3 loci.

[0085] **Example 4**: Pre-processing of sequence data

[0086] Sequencer data processing involves a series of steps to remove errors in the primary sequence of each read, and to compress the data. First, a complexity filter removes approximately 20% of the sequences which are misreads from the sequencer. Then, sequences were required to have a minimum of a six base match to both one of the thirteen J-regions and one of 54 V-regions. Applying the filter to the control lane containing phage sequence, on average only one sequence in 7-8 million passed these steps without false positives. Finally, a nearest neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences, in order to remove both PCR error and sequencing error (see Table 10).

[0087] **Example 5**: Estimating relative CDR3 sequence abundance in PCR pools and blood samples

[0088] After collapsing the data, the underlying distribution of T-cell sequences in the blood reconstructing were derived from the sequence data. The procedure used three steps; 1) flow sorting T-cells drawn from peripheral blood, 2) PCR amplification, and 3) sequencing. Analyzing the data, the ratio of sequences in the PCR product must be derived working backward from the sequence data before estimating the true distribution of clonotypes in the blood.

[0089] For each sequence observed a given number of times in the data herein, the probability that that sequence was sampled from a particular size PCR pool is estimated. Because the CDR3 regions sequenced are sampled randomly from a massive pool of PCR products, the number of observations for each sequence are drawn from Poisson distributions. The Poisson parameters are quantized according to the number of T cell genomes that provided the template for PCR. A simple Poisson mixture model both estimates these parameters and places a pairwise probability

for each sequence being drawn from each distribution. This is an expectation maximization method which reconstructs the abundances of each sequence that was drawn from the blood.

[0090] **Example 6**: Unseen species model for estimation of true diversity

[0091] A mixture model can reconstruct the frequency of each TCRβ CDR3 species drawn from the blood, but the larger question is how many unique CDR3 species were present in the donor? This is a fundamental question that needs to be answered as the available sample is limited in each donor, and will be more important in the future as these techniques are extrapolated to the smaller volumes of blood that can reasonably be drawn from patients undergoing treatment.

[0092] The mathematical solution provides that for a total number of TCRβ "species" or clonotypes, S, a sequencing experiment observes x_s copies of sequence s. For all of the unobserved clonotypes, x_s equals 0, and each TCR clonotype is "captured" in a blood draw according to a Poisson process with parameter λ_s. The number of T cell genomes sequenced in the first measurement 1, and in the second measurement. Since there are a large number of unique sequences, an integral will represent the sum. If $G(\lambda)$ is the empirical distribution function of the parameters $\lambda_1, \ldots, \lambda_S$, and n_x is the number of clonotypes sequenced exactly x times, then

$$E(n_x) = S\int_0^\infty \left(\frac{e^{-\lambda}\lambda^x}{x!}\right) dG(\lambda).$$

[0093] The value Δ(t) is the number of *new* clonotypes observed in the second sequencing experiment.

$$\Delta(t) = \sum_x E(n_x)_{\exp 1+\exp 2} - \sum_x E(n_x)_{\exp 1} = S\int_0^\infty e^{-\lambda}\left(1-e^{-\lambda t}\right) dG(\lambda)$$

[0094] Taylor expansion of $1\text{-}e^{-\lambda t}$ gives $\Delta(t) = E(x_1)t\text{-}E(x_2)t^2+E(x_3)t^3\text{-}\ldots$, which can be approximated by replacing the expectations ($E(n_x)$) with the observed numbers in the first measurement. Using in the numbers observed in the first measurement, this formula predicts that $1.6*10^5$ new unique sequences should be observed in the second measurement. The actual value of the second measurement was $1.8*10^5$ new TCRβ sequences, which implies that the prediction provided a valid lower bound on total diversity. An Euler's transformation was used to regularize $\Delta(t)$ to produce a lower bound for $\Delta(\infty)$.

[0095] **Example 7**: Error correction and bias assessment

[0096] Sequence error in the primary sequence data derives primarily from two sources: (1) nucleotide misincorporation that occurs during the amplification by PCR of TCRβ CDR3 template sequences, and (2) errors in base calls introduced during sequencing of the PCR-amplified library of CDR3 sequences. The large quantity of data allows us to implement a straightforward error correcting code to correct most of the errors in the primary sequence data that are attributable to these two sources. After error correction, the number of unique, in-frame CDR3 sequences and the number of observations of each unique sequence were tabulated for each of the four flow-sorted T cell populations from the two donors. The relative frequency distribution of CDR3 sequences in the four flow cytometrically-defined populations demonstrated that antigen-experienced $CD45RO^+$ populations contained significantly more unique CDR3 sequences with high relative frequency than the $CD45RO^-$ populations. Frequency histograms of TCRβ CDR3 sequences observed in four different T cell subsets distinguished by expression of CD4, CD8, and CD45RO and present in blood showed that ten unique sequences were each observed 200 times in the $CD4^+CD45RO^+$ (antigen-experienced) T cell sample, which was more than twice as frequent as that observed in the $CD4^+CD45RO^-$ populations.

[0097] The use of a PCR step to amplify the TCRβ CDR3 regions prior to sequencing could potentially introduce a systematic bias in the inferred relative abundance of the sequences, due to differences in the efficiency of PCR amplification of CDR3 regions utilizing different Vβ and Jβ gene segments. To estimate the magnitude of any such bias, the TCRβ CDR3 regions from a sample of approximately 30,000 unique $CD4^+CD45RO^+$ T lymphocyte genomes were amplified through 25 cycles of PCR, at which point the PCR product was split in half. Half was set aside, and the other half of the PCR product was amplified for an additional 15 cycles of PCR, for a total of 40 cycles of amplification. The PCR products amplified through 25 and 40 cycles were then sequenced and compared. Over 95% of the 25 cycle sequences were also found in the 40-cycle sample: a linear correlation is observed when comparing the frequency of sequences between these samples. For sequences observed a given number of times in the 25 cycle lane, a combination of PCR bias and sampling variance accounts for the variance around the mean of the number of observations at 40 cycles. Conservatively attributing the mean variation about the line (1.5-fold) entirely to PCR bias, each cycle of PCR amplification potentially introduces a bias of average magnitude $1.5^{1/15} = 1.027$. Thus, the 25 cycles of PCR introduces a total bias of

average magnitude $1.027^{25} = 1.95$ in the inferred relative abundance of distinct CDR3 region sequences.

[0098] **Example 8**: Jβ gene segment usage

[0099] The CDR3 region in each TCR β chain includes sequence derived from one of the thirteen J_β gene segments. Analysis of the CDR3 sequences in the four different T cell populations from the two donors demonstrated that the fraction of total sequences which incorporated sequences derived from the thirteen different J_β gene segments varied more than 20-fold. Jβ utilization among four different T flow cytometrically-defined T cells from a single donor is was relatively constant within a given donor. Moreover, the J_β usage patterns observed in two donors, which were inferred from analysis of genomic DNA from T cells sequenced using the GA, are qualitatively similar to those observed in T cells from umbilical cord blood and from healthy adult donors, both of which were inferred from analysis of cDNA from T cells sequenced using exhaustive capillary-based techniques.

[0100] **Example 9**: Nucleotide insertion bias

[0101] Much of the diversity at the CDR3 junctions in TCR α and β chains is created by non-templated nucleotide insertions by the enzyme Terminal Deoxynucloetidyl Transferase (TdT). However, in vivo, selection plays a significant role in shaping the TCR repertoire giving rise to unpredictability. The TdT nucleotide insertion frequencies, independent of selection, were calculated using out of frame TCR sequences. These sequences are non-functional rearrangements that are carried on one allele in T cells where the second allele has a functional rearrangement. The mono-nucleotide insertion bias of TdT favors C and G (Table 11).

Table 11: Mono-nucleotide bias in out of frame data

	A	C	G	T
Lane 1	0.24	0.294	0.247	0.216
Lane 2	0.247	0.284	0.256	0.211
Lane 3	0.25	0.27	0.268	0.209
Lane 4	0.255	0.293	0.24	0.21

[0102] Similar nucleotide frequencies are observed in the in frame sequences (Table 12).

Table 12: Mono-nucleotide bias in in-frame data

	A	C	G	T
Lane 1	0.21	0.285	0.275	0.228
Lane 2	0.216	0.281	0.266	0.235
Lane 3	0.222	0.266	0.288	0.221
Lane 4	0.206	0.294	0.228	0.27

[0103] The N regions from the out of frame TCR sequences were used to measure the di-nucleotide bias. To isolate the marginal contribution of a di-nucleotide bias, the di-nucleotide frequencies was divided by the mononucleotide frequencies of each of the two bases. The measure is

$$m = \frac{f(n_1 n_2)}{f(n_1) f(n_2)}.$$

[0104] The matrix for m is found in Table 13.

Table 13: Di-nucleotide odd ratios for out of frame data

	A	C	G	T
A	1.198	0.938	0.945	0.919
C	0.988	1.172	0.88	0.931
G	0.993	0.701	1.352	0.964
T	0.784	1.232	0.767	1.23

[0105] Many of the dinucleotides are under or over represented. As an example, the odds of finding a GG pair are very high. Since the codons GGN translate to glycine, many glycines are expected in the CDR3 regions.

[0106] **Example 10**: Amino Acid distributions in the CDR3 regions

[0107] The distribution of amino acids in the CDR3 regions of TCRβ chains are shaped by the germline sequences for V, D, and J regions, the insertion bias of TdT, and selection. The distribution of amino acids in this region for the four different T cell sub-compartments is very similar between different cell subtypes. Separating the sequences into β chains of fixed length, a position dependent distribution among amino acids, which are grouped by the six chemical properties: small, special, and large hydrophobic, neutral polar, acidic and basic. The

distributions are virtually identical except for the CD8+ antigen experienced T cells, which have a higher proportion of acidic bases, particularly at position 5.

[0108] Of particular interest is the comparison between $CD8^+$ and $CD4^+$ TCR sequences as they bind to peptides presented by class I and class II HLA molecules, respectively. The $CD8^+$ antigen experienced T cells have a few positions with a higher proportion of acidic amino acids. This could be do binding with a basic residue found on HLA Class I molecules, but not on Class II.

[0109] **Example 11**: TCR β chains with identical amino acid sequences found in different people

[0110] The TCR β chain sequences were translated to amino acids and then compared pairwise between the two donors. Many thousands of exact sequence matches were observed. For example, comparing the $CD4^+$ $CD45RO^-$ sub-compartments, approximately 8,000 of the 250,000 unique amino acid sequences from donor 1 were exact matches to donor 2. Many of these matching sequences at the amino acid level have multiple nucleotide differences at third codon positions. Following the example mentioned above, 1,500/8,000 identical amino acid matches had >5 nucleotide mismatches. Between any two T cell sub-types, 4-5% of the unique TCRβ sequences were found to have identical amino acid matches.

[0111] Two possibilities were examined: that 1) selection during TCR development is producing these common sequences and 2) the large bias in nucleotide insertion frequency by TdT creates similar nucleotide sequences. The in-frame pairwise matches were compared to the out-of-frame pairwise matches (see Examples 1-4, above). Changing frames preserved all of the features of the genetic code and so the same number of matches should be found if the sequence bias was responsible for the entire observation. However, almost twice as many in-frame matches as out-of-frame matches were found, suggesting that selection at the protein level is playing a significant role.

[0112] To confirm this finding of thousands of identical TCR β chain amino acid sequences, two donors were compared with respect to the $CD8^+$ $CD62L^+$ $CD45RA^+$ (naïve-like) TCRs from a third donor, a 44 year old CMV^+ Caucasian female. Identical pairwise matches of many thousands of sequences at the amino acid level between the third donor and each of the original two donors were found. In contrast, 460 sequences were shared between all three donors. The large variation in total number of unique sequences between the donors is a product of the starting material and variations in loading onto the sequencer, and is not representative of a variation in true diversity in the blood of the donors.

[0113] **Example 12**: Higher frequency clonotypes are closer to germline

[0114] The variation in copy number between different sequences within every T cell sub-compartment ranged by a factor of over 10,000-fold. The only property that correlated with copy number was (the number of insertions plus the number of deletions), which inversely correlated. Results of the analysis showed that deletions play a smaller role than insertions in the inverse correlation with copy number.

[0115] Sequences with less insertions and deletions have receptor sequences closer to germ line. One possibility for the increased number of sequences closer to germ line is that they are the created multiple times during T cell development. Since germ line sequences are shared between people, shared TCRβ chains are likely created by TCRs with a small number of insertions and deletions.

[0116] **Example 13**: "Spectratype" analysis of TCRβ CDR3 sequences by V gene segment utilization and CDR3 length

[0117] TCR diversity has commonly been assessed using the technique of TCR spectratyping, an RT-PCR-based technique that does not assess TCR CDR3 diversity at the sequence level, but rather evaluates the diversity of TCRα or TCRβ CDR3 lengths expressed as mRNA in subsets of $\alpha\beta$ T cells that use the same V_α or V_β gene segment. The spectratypes of polyclonal T cell populations with diverse repertoires of TCR CDR3 sequences, such as are seen in umbilical cord blood or in peripheral blood of healthy young adults typically contain CDR3 sequences of 8-10 different lengths that are multiples of three nucleotides, reflecting the selection for in-frame transcripts. Spectratyping also provides roughly quantitative information about the relative frequency of CDR3 sequences with each specific length. To assess whether direct sequencing of TCRβ CDR3 regions from T cell genomic DNA using the sequencer could faithfully capture all of the CDR3 length diversity that is identified by spectratyping, "virtual" TCRβ spectratypes (see Examples above) were generated from the sequence data and compared with TCRβ spectratypes generated using conventional PCR techniques. The virtual spectratypes contained all of the CDR3 length and relative frequency information present in the conventional spectratypes. Direct TCRβ CDR3 sequencing captures all of the TCR diversity information present in a conventional spectratype. A comparison of standard TCRβ spectratype data and calculated TCRβ CDR3 length distributions for sequences utilizing representative TCR Vβ gene segments and present in $CD4^+CD45RO^+$ cells from donor 1. Reducing the information contained in the sequence data to a

frequency histogram of the unique CDR3 sequences with different lengths within each Vβ family readily reproduces all of the information contained in the spectratype data. In addition, the virtual spectratypes revealed the presence within each V_β family of rare CDR3 sequences with both very short and very long CDR3 lengths that were not detected by conventional PCR-based spectratyping.

[0118] **Example 14**: Estimation of total CDR3 sequence diversity

[0119] After error correction, the number of unique CDR3 sequences observed in each lane of the sequencer flow cell routinely exceeded 1×10^5. Given that the PCR products sequenced in each lane were necessarily derived from a small fraction of the T cell genomes present in each of the two donors, the total number of unique TCRβ CDR3 sequences in the entire T cell repertoire of each individual is likely to be far higher. Estimating the number of unique sequences in the entire repertoire, therefore, requires an estimate of the number of additional unique CDR3 sequences that exist in the blood but were not observed in the sample. The estimation of total species diversity in a large, complex population using measurements of the species diversity present in a finite sample has historically been called the "unseen species problem" (see Examples above). The solution starts with determining the number of new species, or TCRβ CDR3 sequences, that are observed if the experiment is repeated, i.e., if the sequencing is repeated on an identical sample of peripheral blood T cells, e.g., an identically prepared library of TCRβ CDR3 PCR products in a different lane of the sequencer flow cell and counting the number of new CDR3 sequences. For $CD8^+CD45RO^-$ cells from donor 2, the predicted and observed number of new CDR3 sequences in a second lane are within 5% (see Examples above), suggesting that this analytic solution can, in fact, be used to estimate the total number of unique TCRβ CDR3 sequences in the entire repertoire.

[0120] The resulting estimates of the total number of unique TCRβ CDR3 sequences in the four flow cytometrically-defined T cell compartments are shown in Table 14.

Table 14: TCR repertoire diversity

Donor	CD8	CD4	CD45RO	Diversity
1	+	-	+	$6.3*10^5$
	+	-	-	$1.24*10^6$
	-	+	+	$8.2*10^5$
	-	+	-	$1.28*10^6$
	Total T cell diversity			$3.97*10^6$
2	+	-	+	$4.4*10^5$
	+	-	-	$9.7*10^5$
	-	+	+	$8.7*10^5$
	-	+	-	$1.03*10^6$
	Total T cell diversity			$3.31*10^6$

[0121] Of note, the total TCRβ diversity in these populations is between 3-4 million unique sequences in the peripheral blood. Surprisingly, the CD45RO$^+$, or antigen-experienced, compartment constitutes approximately 1.5 million of these sequences. This is at least an order of magnitude larger than expected. This discrepancy is likely attributable to the large number of these sequences observed at low relative frequency, which could only be detected through deep sequencing. The estimated TCRβ CDR3 repertoire sizes of each compartment in the two donors are within 20% of each other.

[0122] The results herein demonstrate that the realized TCRβ receptor diversity is at least five-fold higher than previous estimates (~$4*10^6$ distinct CDR3 sequences), and, in particular, suggest far greater TCRβ diversity among CD45RO$^+$ antigen-experienced $\alpha\beta$ T cells than has previously been reported (~$1.5*10^6$ distinct CDR3 sequences). However, bioinformatic analysis of the TCR sequence data shows strong biases in the mono- and di- nucleotide content, implying that the utilized TCR sequences are sampled from a distribution much smaller than the theoretical size. With the large diversity of TCRβ chains in each person sampled from a severely constrict space of sequences, overlap of the TCR sequence pools can be expected between each person. In fact, the results showed about 5% of CD8$^+$ naïve TCRβ chains with exact amino acid matches are shared between each pair of three different individuals. As the TCRα pool has been previously measured to be substantially smaller than the theoretical TCRβ diversity, these results show that hundreds to thousands of truly public $\alpha\beta$ TCRs can be found.

What is Claimed:

1. A composition comprising:
   (a) a multiplicity of V-segment primers, wherein each primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
   (b) a multiplicity of J-segment primers, wherein each primer comprises a sequence that is complementary to a J segment;
   wherein the V segment and J-segment primers permit amplification of a TCR or IG CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of the TCR or IG genes.

2. The composition of claim 1, wherein each V-segment primer comprises a sequence that is complementary to a single Vγ segment or a family of similar Vγ segments, and each J segment primer comprises a sequence that is complementary to a Jγ segment, and wherein V segment and J-segment primers permit amplification of a TCRγ CDR3 region.

3. The composition of claim 1, wherein each V-segment primer comprises a sequence that is complementary to a single Vδ segment or a family of similar Vδ segments, and each J segment primer comprises a sequence that is complementary to a Jδ segment, and wherein V segment and J-segment primers permit amplification of a TCRδ CDR3 region.

4. The composition of claim 1, wherein each V-segment primer comprises a sequence that is complementary to a single Vα segment or a family of similar Vα segments, and each J segment primer comprises a sequence that is complementary to a Jα segment, and wherein V segment and J-segment primers permit amplification of a TCRα CDR3 region.

5. The composition of claim 1, wherein each V-segment primer comprises a sequence that is complementary to a single Vβ segment or a family of similar Vβ segments, and each J segment primer comprises a sequence that is complementary to a Jβ segment, and wherein V segment and J-segment primers permit amplification of a TCRβ CDR3 region.

6. The composition of claim 1, wherein the V segment have similar annealing strength.

7. The composition of claim 1, wherein all J segment primers anneal to the same conserved framework region motif.

8. The composition of claim 1, wherein the amplified DNA molecule starts from said conserved motif and diagnostically identifies the J segment and includes the junction and into the V segment.

9. The composition of claim 1, further comprising a set of sequencing oligonucleotides, wherein the sequencing oligonucleotides hybridize to a regions within the amplified DNA molecules.

10. The composition of claim 1, wherein the amplified DNA spans a V-D-J junction.

11. The composition of claim 1, wherein the V-segment or J-segment are selected to contain a sequence error-correction by merger of closely related sequences.

12. The composition of claim 1, further comprising a universal C segment primer for generating cDNA from mRNA.

13. The composition of claim 5, wherein the V segment primer is anchored at position -43 in the Vβ segment relative to the recombination signal sequence (RSS).

14. The composition of claim 5, wherein the multiplicity of V segment primers consist of at least 14 primers specific to 14 different Vβ genes.

15. The composition of claim 5, wherein the V segment primers have sequences that are selected from the group consisting of SEQ ID NOS:1-45.

16. The composition of claim 5, wherein the V segment primers have sequences that are selected from the group consisting of SEQ ID NOS:58-102.

17. The composition of claim 5, wherein there is a V segment primer for each Vβ segment or family of Vβ segments.

18. The composition of claim 5, wherein the primers do not cross an intron/exon boundary.

19. The composition of claim 5, wherein the J segment primers hybridize with a conserved element of the Jβ segment, and have similar annealing strength.

20. The composition of claim 5, wherein the multiplicity of J segment primers consist of at least five primers specific to five different Jβ genes.

21. The composition of claim 5, wherein the J segment primers have sequences that are selected from the group consisting of SEQ ID NOS:46-57 and 483.

22. The composition of claim 5, wherein the J segment primers have sequences that are selected from the group consisting of SEQ ID NOS:103-113, 468 and 484.

23. The composition of claim 5, wherein there is a J segment primer for each Jβ segment.

24. The composition of claim 5, wherein the amplified Jβ gene segments each have a unique four base tag at positions +11 through +14 downstream of the RSS site.

25. The composition of claim 24, wherein the sequencing oligonucleotides hybridize adjacent to a four base tag within the amplified Jβ gene segments at positions +11 through +14 downstream of the RSS site.

26. The composition of claim 24, wherein the sequencing oligonucleotides are selected from the group consisting of SEG ID NOS:470-482.

27. A composition comprising:
    (a) a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
    (b) a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;
    wherein the V segment and J segment primers permit amplification of antibody heavy chain (IGH) V_H region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of antibody heavy chain genes.

28. A composition comprising:
    (a) a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and

(b) a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;

wherein the V segment and J segment primers permit amplification of antibody light chain (IGL) V_L region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of antibody light chain genes.

29. A method comprising:

(a) selecting a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and

(b) selecting a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;

(c) combining the V segment and J segment primers with a sample of genomic DNA to permit amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules sufficient to quantify the diversity of the TCR genes.

30. The method of claim 29, wherein each V segment primer comprises a sequence that is complementary to a single Vβ segment or a family of Vβ segments, and each J segment primer comprises a sequence that is complementary to a Jβ segment; and wherein combining the V segment and J segment primers with a sample of genomic DNA permits amplification of a TCRB CDR3 region by a multiplex polymerase chain reaction (PCR) and produces a multiplicity of amplified DNA molecules.

31. The method of claim 30, further comprising a step of sequencing the amplified DNA molecules.

32. The method of claim 31, wherein the sequencing step utilizes a set of sequencing oligonucleotides that hybridize to a defined region within the amplified DNA molecules.

33. The method of claim 32, further comprising a step of calculating the total diversity of TCRβ CDR3 sequences among the amplified DNA molecules.

34. The method of claim 33, wherein the method shows that the total diversity of a normal human subject is greater than $1*10^6$ sequences.

35. The method of claim 33, wherein the method shows that the total diversity of a normal human subject is greater than $2*10^6$ sequences.

36. The method of claim 33, wherein the method shows that the total diversity of a normal human subject is greater than $3*10^6$ sequences.

37. A method of diagnosing immunodeficiency in a human patient, comprising measuring the diversity of TCR CDR3 sequences of the patient, and comparing the diversity of the subject to the diversity obtained from a normal subject.

38. The method of claim 37, wherein measuring the diversity of TCR sequences comprises the steps of:
    (a) selecting a multiplicity of V segment primers, wherein each V segment primer comprises a sequence that is complementary to a single functional V segment or a small family of V segments; and
    (b) selecting a multiplicity of J segment primers, wherein each J segment primer comprises a sequence that is complementary to a J segment;
    (c) combining the V segment and J segment primers with a sample of genomic DNA to permit amplification of a TCR CDR3 region by a multiplex polymerase chain reaction (PCR) to produce a multiplicity of amplified DNA molecules;
    (d) sequencing the amplified DNA molecules;
    (e) calculating the total diversity of TCR CDR3 sequences among the amplified DNA molecules.

39. The method of claim 38, wherein comparing the diversity is determined by calculating using the following equation:

$$\Delta(t) = \sum_x E(n_x)_{measurement1+2} - \sum_x E(n_x)_{measurement2} = S\int_0^\infty e^{-\lambda}\left(1-e^{-\lambda t}\right)dG(\lambda)$$

wherein $G(\lambda)$ is the empirical distribution function of the parameters $\lambda_1, \ldots, \lambda_S$, n_x is the number of clonotypes sequenced exactly x times, and

$$E(n_x) = S\int_0^\infty \left(\frac{e^{-\lambda}\lambda^x}{x!}\right)dG(\lambda).$$

40. The method of claim 38, wherein the diversity of at least two samples of genomic DNA are compared.

41. The method of claim 40, wherein one sample of genomic DNA is from a patient and the other sample is from a normal subject.

42. The method of claim 40, wherein one sample of genomic DNA is from a patient before a therapeutic treatment and the other sample is from the patient after treatment.

43. The method of claim 40, wherein the two samples of genomic DNA are from the same patient at different times during treatment.

44. The method of claim 40, in which a disease is diagnosed based on the comparison of diversity among the samples of genomic DNA.

45. The method of claim 40, wherein the immunocompetence of a human patient is assessed by the comparison.

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/037477

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MARIANI SARA ET AL: "Comprehensive assessment of the TCRBV repertoire in small T-cell samples by means of an improved and convenient multiplex PCR method" EXPERIMENTAL HEMATOLOGY (NEW YORK), vol. 37, no. 6, 20 May 2009 (2009-05-20), - 20 May 2009 (2009-05-20) pages 728-738, XP002596859 ISSN: 0301-472X the whole document	1-45
X	EP 2 062 982 A1 (IMMUNID [FR]; COMMISSARIAT ENERGIE ATOMIQUE [FR]) 27 May 2009 (2009-05-27) the whole document	1-45

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 September 2010	24/09/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Dolce, Luca

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/037477

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VAN DONGEN J ET AL: "Design and standardization of PCR primers and protocols for detection of clonal immunoglobuline and T-cell receptor gene recombinations is suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936" LEUKEMIA, MACMILLAN PRESS LTD, US, vol. 17, no. 12, 1 December 2000 (2000-12-01), pages 2257-2317, XP008093070 ISSN: 0887-6924 the whole document	1-45
X	MASLANKA KRYSTYNA ET AL: "Molecular analysis of T cell repertoires: Spectratypes generated by multiplex polymerase chain reaction and evaluated by radioactivity or fluorescence" HUMAN IMMUNOLOGY, vol. 44, no. 1, 1995, pages 28-34, XP002596860 ISSN: 0198-8859 the whole document	1-45
X	KIIANITSA KONSTANTIN ET AL: "Development of tools for T cell repertoire analysis (TCRB spectratyping) for the canine model of hematopoietic cell transplantation" BLOOD, vol. 110, no. 11, Part 2, November 2007 (2007-11), page 293B, XP002596861 & 49TH ANNUAL MEETING OF THE AMERICAN-SOCIETY-OF-HEMATOLOGY; ATLANTA, GA, USA; DECEMBER 08 -11, 2007 ISSN: 0006-4971 the whole document	1-45

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2010/037477

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 2062982 A1	27-05-2009	CA 2706667 A1	06-08-2009
		EP 2220255 A2	25-08-2010
		WO 2009095567 A2	06-08-2009