

(51) International Patent Classification:
C12Q 1/68 (2006.01)(21) International Application Number:
PCT/US2011/026373(22) International Filing Date:
25 February 2011 (25.02.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/308,261 25 February 2010 (25.02.2010) 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): **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). **ROBINS, Harlan, S.** [US/US]; 4418 Latona Avenue N.e., Seattle, WA 98105 (US).(74) Agents: **SAWISLAK, Deborah A.** et al.; Woodcock Washburn LLP, 2929 Arch Street, Cira Centre, 12th Floor, 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, RS, 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))

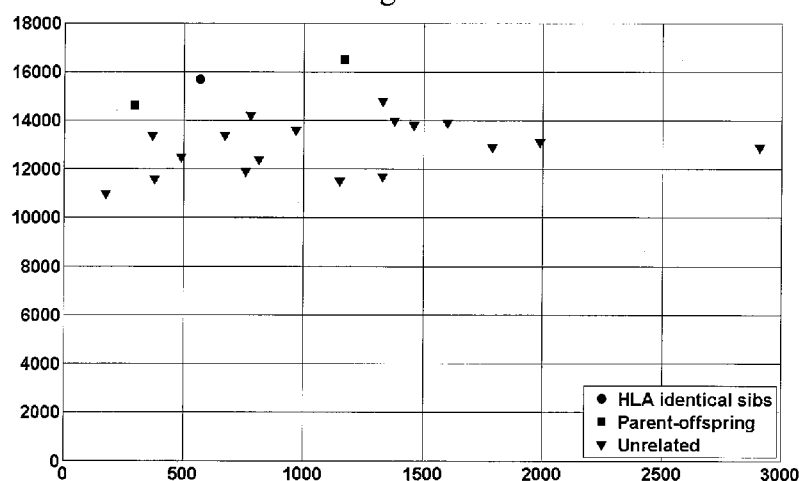
Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

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

(54) Title: USE OF TCR CLONOTYPES AS BIOMARKERS FOR DISEASE

Figure 1



(57) Abstract: A method of detecting disease or risk of disease in subject using a biomarker or panel of biomarkers identified from exposure-specific clonotype sequences is described. The method for identifying the exposure-specific clonotype sequences from a subset of B cells or T cells from an immune cell sample is also described.

USE OF TCR CLONOTYPES AS BIOMARKERS FOR DISEASE

REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 61/308,261 filed February 25, 2010, and is hereby incorporated by reference in its entirety.

BACKGROUND

Identifying exposure to specific environmental agents, ranging from bacteria to chemical carcinogens is important for detecting, treating and monitoring a variety of human diseases. As part of the immune response to pathogens, B cells and T cells in the adaptive immune system form long lasting memory responses to protect against future exposures which can be measured. The response specific to antigens is based on the highly polymorphic receptors encoded by B cells (immunoglobulins (IG)) and T cells (T cell receptors (TCR)). TCRs are heterodimeric proteins consisting of either an α chain and a β chain or a γ chain and a δ chain. A similar structure in B cell receptors is present where IGs are also a heterodimer, consisting of one light and one heavy chain.

Diversity in lymphocyte antigen receptors is generated by somatic rearrangement of TCR and IG genes, and is concentrated within the third complementarity-determining region (CDR3) of each chain of the receptor heterodimer. In TCRB and IGH, the nucleotide sequences that encode the CDR3 regions are noncontiguous variable (V), diversity (D), and joining (J) region gene segments, while only V and J segments are used for the TCRA and IGL chains. The existence of multiple V, D, and J gene segments in germline DNA permits substantial combinatorial diversity in receptor composition. Receptor diversity is further augmented by the deletion of nucleotides adjacent to the recombination signal sequences (RSS) of the V, D, and J segments, and template-independent insertion of nucleotides at the V-D, D-J, and V-J junctions.

Infectious disease diagnostics presently rely upon some similarity of symptoms as a basis to screen for exposure to single pathogens or small panels of pathogens. Furthermore, in the past discoveries of association between exposure and disease state have required prior hypotheses as to the identity of the exposure. The ability to screen for environmental exposures that correlate with disease states without pre-specifying the exposure would be invaluable to public health efforts by improving measurements of environmental exposures. The sequencing depth of

capillary-based technologies does not make a comprehensive comparison of memory IG and TCR variable sequences between individuals feasible. The present invention provides methods and compositions for the identification of sequences specific to pathogenic exposures and other uses that should be apparent to those skilled in the art from the teachings herein.

SUMMARY

In one aspect, the present invention provides a method for identifying a biomarker for a disease comprising a) providing isolated polynucleotide sequences from immune cells from a group of patients with the disease; b) performing a nucleotide amplification reaction to produce a set of clonotype sequences; c) identifying clonotype sequences enriched within the group of patients; d) providing isolated polynucleotide sequences from immune cells from a group of normal subjects without the disease, and amplifying the polynucleotides; e) removing sequences present in the normal subject group, which are obtained in step d), from the exposure-specific clonotype sequences, which are obtained in step c).

Included in the method are certain embodiments as described herein. In one embodiment, the polynucleotide sequences are rearranged genomic sequences. In another embodiment, the nucleotide amplification reaction comprises (i) a multiplicity of V-segment primers, where each primer comprises a sequence that is complementary to a single functional V-segment or a small family of V-segments; and (ii) a multiplicity of J-segment primers, where each primer comprises a sequence that is complementary to a J-segment; and where the V-segment and J-segment primers amplify a TCR CDR3 region. In another embodiment, the V-segments are selected from the group consisting of V α , V β , V γ and V δ , and the J-segments are selected from the group consisting of J α , J β , J γ and J δ . In other embodiments, the immune cell samples are selected from the group consisting of CD45RO⁺, CD45RA^{int/neg} CD8⁺ T cells and CD45RO⁻ CD45RA^{hi} CD62L^{hi} CD8⁺ T cells. In certain embodiments, the T cells share one or more HLA alleles. In another embodiment, (i) 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 (ii) a multiplicity of J primers, wherein each primer comprises a sequence that is complementary to a J-segment; wherein the V-segment and J-segment primers amplify an IGH, IGL or IGK CDR3 region. In another embodiment, the V-segment primers (forward primers) are anchored at a position between 40 and 60 base pairs 5' of the recombination signal sequence (RSS), within the V segment. Another embodiment can include

where the J-segment primers (reverse primers) are at a position about 30 base pair 3' of the J gene RSS site. Certain other embodiments may include where the exposure-specific clonotype sequences have an insertion of less than six nucleotides.

In other embodiments of the present invention, certain diseases are of importance and include diseases selected from the group consisting of an autoimmune disease, an inflammatory disease, an immune deficiency, a bacterial infection, a viral infection, a fungal infection, or a parasitic infection.

Another embodiment of the present invention provides a method comprising an additional step prior to step (a) of isolating polynucleotide sequences from the immune cell samples. Further embodiments include polynucleotide sequences isolated from an immune cell sample, where the sample is a tissue comprising hematopoietic lineage cells.

The present invention also provides biomarkers produced by the methods described herein. In certain embodiments, the biomarker is polypeptide encoded by the exposure-specific clonotype sequence. In another embodiment, the polypeptide biomarker is an antibody.

In another aspect, the present invention provides a diagnostic kit for detecting a disease or risk for a disease comprising one or more biomarkers described herein. In certain embodiments, the biomarker is selected from the group consisting of a polynucleotide biomarker, a labeled polypeptide biomarker, and an antibody biomarker. In embodiments of the present invention, the biomarker detects a disease or risk of a disease selected from the group consisting of an autoimmune disease, an inflammatory disease, an immune deficiency, a bacterial infection, a viral infection, a fungal infection, a parasite infection, and a prenatal disease.

In another aspect, the present invention provides a method for detecting disease or a risk for disease in a subject comprising a method for detecting a disease or a risk of a disease in subject in a subject comprising: a) selecting a biomarker for the disease; b) detecting the presence of the biomarker in the polypeptide sequences of the subject. In other embodiments, the method provides further comprises determining that the presence of the biomarker in the polypeptide sequences of the subject is indicative of the disease or the risk of the disease.

In certain embodiments, the selection uses a panel of biomarkers, and (c) comprises determining if the frequency of the exposure-specific clonotype sequence is indicative of disease or risk of disease. In other embodiments, the clonotype sequences are genomic sequences.

It is another aspect of the present invention to provide a method for detecting disease or a risk of disease in subject in a subject comprising (a) providing isolated clonotype sequences from

the subject's immune cell sample; (b) detecting the presence of an exposure-specific clonotype sequence using the claimed biomarkers described herein; and (c) determining if the frequency of the exposure-specific clonotype sequence from the subject is at a frequency indicative of disease or a risk of disease. In certain embodiments, the method comprises using a panel of biomarkers and (c) comprises determining if the frequency of the exposure-specific clonotype sequence is indicative of disease or risk of disease.

In another aspect, the present invention provides a method for selecting a biomarker for diabetes or a risk for Type I diabetes comprising selecting a biomarker according to the method described herein, where the exposure-specific clonotype sequences are selected from the group consisting of SEQ ID. NOS: 75 to 763.

In another aspect, the present invention provides a method for detecting diabetes Type I or a risk for diabetes Type I in a subject comprising (a) selecting a disease-specific biomarker from SEQ ID. NOS: 75 to 763, where the biomarker consists of a disease specific clonotype sequence or a panel of such sequences; (b) measuring the frequency of the clonotype sequence(s) in the subject; and (c) determining if the frequency of the diagnostic clonotype sequences in the subject is more consistent with frequencies observed in diseases cases or controls.

In another aspect, the present invention provides a method for detecting multiple sclerosis or a risk for multiple sclerosis in a subject comprising selecting a biomarker according to the methods described herein, where the exposure-specific clonotype sequences are selected from the group consisting of SEQ ID. NOS: 764 to 1166.

In another aspect, the present invention provides a method for detecting multiple sclerosis or a risk for multiple sclerosis in a subject comprising (a) selecting a disease-specific biomarker from SEQ ID. NOS: 764 to 1166, where the biomarker consists of a disease specific clonotype sequence or a panel of such sequences; (b) measuring the frequency of the clonotype sequence(s) in the subject; and (c) determining if the frequency of the diagnostic clonotype sequences in the subject is more consistent with frequencies observed in diseases cases or controls.

In another aspect, the present invention provides a method for selecting a biomarker for diabetes or a risk for Epstein Barr Virus comprising selecting a biomarker according to the method described herein, wherein the exposure-specific clonotype sequences are selected from the group consisting of SEQ ID. NOS: 71 to 74 and 1167.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1: Pairwise shared CD8+ TCR beta amino acid sequences. The T cell repertoires from seven healthy people were compared. Each point on the graph represents a pairwise comparison between two of the seven donors for a total of 21 comparisons. The Y-axis value is number of matching TCR beta chains from the naive T cells found in the blood. The X-axis value is the number of matching TCR beta chains from the memory T cells in the blood.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

The following definitions are intended to assist in understanding the present invention:

“Nucleotide” or “nucleotide sequence” as used herein refers to polynucleotides, such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), oligonucleotides, fragments generated by the polymerase chain reaction (PCR), and fragments generated by any of ligation, scission, endonuclease action, and exonuclease action. Nucleic acid molecules can be composed of monomers that are naturally occurring nucleotides (such as DNA and RNA), or analogs of naturally-occurring nucleotides (e.g., a-enantiomeric forms of naturally-occurring nucleotides), or a combination of both, unless otherwise specified. Those skilled in the art will recognize that considerable sequence variation is possible among nucleotide molecules encoding for amino acids based on the degeneracy of the genetic code. Modified nucleotides can have alterations in sugar moieties and/or in pyrimidine or purine base moieties.

The term "complement of a nucleic acid molecule" refers to a nucleic acid molecule having a complementary nucleotide sequence and reverse orientation as compared to a reference nucleotide sequence.

As used herein, a "primer" defines an oligonucleotide which is capable of annealing to (hybridizing with) a target sequence, thereby creating a double stranded region which can serve as an initiation point for polynucleotide synthesis under suitable conditions.

The term “nucleotide amplification reaction” refers to any suitable procedure that amplifies a specific region of polynucleotides (target) using primers. See generally Kwoh *et al.*, *Am. Biotechnol. Lab.* 8:14 (1990); Kwoh *et al.*, *Proc. Natl. Acad. Sci. USA* 86, 1173-1177 (1989); Lizardi *et al.*, *BioTechnology* 6:1197-1202 (1988); Malek *et al.*, *Methods Mol. Biol.*, 28:253-260 (1994); and Sambrook *et al.*, *"Molecular Cloning: A laboratory Manual"* (1989).

A "polypeptide" is a polymer of amino acid residues joined by peptide bonds. As used herein refers to any two or more joined amino acid residues and includes peptides and proteins whether produced naturally or synthetically, unless specifically stated otherwise.

“Antigen” as used herein denotes a molecule that binds antibodies. As used herein antigen also includes molecules that can illicit an adaptive immune response also referred to as “immunogen”.

“Immune cells” as used herein denotes cells derived from a pluripotent hematopoietic stem cell and differentiated from a common lymphoid stem cell into cells of the lymphoid cell lineage. Lymphoid lineage cells give rise B cells and T cells and are located in the lymphoid organs and tissues which include bone marrow, blood, thymus, spleen, lymph nodes and mucosal lymphoid tissues.

“B cell” as used herein denotes a lymphocyte derived from a B cell progenitor in bone marrow. The receptor on the B cell is referred to as a B cell receptor or “immunoglobulin” (IG) and consists of an effector region and a two component homodimer, each having two heterodimeric chains with an immunoglobulin heavy (IG_H) and immunoglobulin light (IG_L). The IG_H is organized into five classes, H α , H δ , H ϵ , H γ , and H μ , and the IG_L is organized into two classes, L κ and L λ . Naïve B cells have not encountered specific antigens and are naive when leaving the bone marrow. Memory B cells upon activation by antigen produce antibodies specific to that antigen.

“T cell” as used herein denotes a lymphocyte that is maintained in the thymus and has either α : β or γ : δ heterodimeric receptor. There are V α , V β , V γ and V δ , J α , J β , J γ and J δ , and D β and D δ loci. Naïve T cells have not encountered specific antigens and T cells are naive when leaving the thymus. Naive T cells are identified as CD45RO⁻, CD45RA^{hi}, CD62L⁺. Memory T cells mediate immunological memory to respond rapidly on re-exposure to the antigen that originally induced their expansion and can be “CD8+” (T helper cells) or “CD4+” (T cytotoxic cells). Memory CD4 T cells are identified as CD4⁺, CD45RO⁺ cells and memory CD8 cells are identified as CD8⁺ CD45RO⁺.

“MHC molecule” as used herein denotes a highly polymorphic glycoprotein encoded by major histocompatibility complex (MHC) class I and II genes, and plays a role in presenting antigen to the T cell.

A "detectable label" is a molecule or atom which can be conjugated to a nucleic acid sequence, polypeptide sequence or antibody moiety to produce a molecule useful for diagnosis. Examples of detectable labels include chelators, photoactive agents, radioisotopes, fluorescent agents, paramagnetic ions, or other marker moieties.

All references cited herein are incorporated by reference in their entirety.

A person's ability to defend against specific environmental agents and pathogens including viruses, bacteria and chemical carcinogens requires recognizing the pathogen and responding in ways that are both immediate and long-lasting. Recognition is a function of both the innate and adaptive systems, but it is the B cells and T cells in the adaptive immune system that form long lasting memory responses to protect against future exposures to the same pathogen. To achieve this, the adaptive immune system employs several ingenious strategies to generate a repertoire of T- and B-cell antigen receptors with sufficient diversity to recognize the universe of potential pathogens. In T cells, which primarily recognize peptide antigens presented by MHC molecules on the surface of specialized antigen presenting cells, most of this receptor diversity is contained within the third complementarity-determining region (CDR3) of the T cell receptor (TCR) α and β chains. In B cell receptors antigenic specificity is determined in the third region of the CDR of the immunoglobulin heavy (IGH) and immunoglobulin light (IGL) peptides, and is not MHC dependent. Two properties of B and T cells that make IGs and TCRs very enticing biomarkers are (1) clonal expansion upon recognition of an antigen and (2) the causal association of the cells with disease resistance (for pathogens) or disease pathology (in autoimmune disease). Upon binding to a target antigen, the immune system amplifies these biomarkers through rapid cell division, allowing these receptors to be readily detectable even if the target is rare (e.g. an early stage tumor). Second, in autoimmune diseases, the immune cells useful as biomarkers might additionally be a causative component of the disease and, therefore, a drug target. For example, dysregulated T cells specifically attacking self tissues are thought to play a major role in multiple sclerosis, Type I diabetes, and rheumatoid arthritis.

Published studies of the human TCR β CDR3 sequence repertoire with high-throughput sequencing technologies have used a variety of pre-sequencing molecular strategies, sequencing platforms, and analytical approaches to assess the repertoire realized in single individuals or a pool of 550 individuals ((Robins, *et al.*, *Blood* 114, 4099-4107 (2009); Venturi, *et al.*, *Nat Rev Immunol* 8, 231-238 (2008); Wang, *et al.*, *Proc Natl Acad Sci U S A* 107, 1518-1523 (2010)). Due to the different molecular strategies employed in the three studies, the number of unique TCR β CDR3 nucleotide sequences observed varied widely, from a low of ~34,000 – derived from analysis of 40.5 million primary sequence reads (Freeman, *et al.* *Genome Res* 19, 1817-1824 (2009)) – to more than 500,000 unique sequences (Robins, *supra*). The present invention is based on assessing the TCR β CDR3 repertoires realized in the naïve and memory CD8+ compartments of multiple donors, and for the first time reading and comparing approximately 3 million unique CDR3 sequences from ~40 million primary sequences from the repertoires of

these different individuals. These unique CDR3 sequences identified from TCR or IG are defined as "clonotype sequences". The results unexpectedly showed half of all TCR β in each individual were derived from a small and specific subset of all possible TCR β , characterized by less than about five inserted nucleotides, and biased toward specific VJ pair usage. Moreover, a subset of TCR β from the donors was in response to the same pathogen, Epstein-Barr Virus. The frequency of these T cells where donors were exposed to the Epstein Barr Virus was small enough to identify clonotype sequences as biomarkers reflecting Epstein Barr Virus infection. The clonotype sequences associated with a single disease or pathogen within an individual are defined as "exposure-specific clonotype sequences". Exposure-specific clonotype sequences observed in common across multiple individuals are defined as "public antigen receptors", and the cells carrying these receptors are "public T cells" or "public B cells". The present specification discloses that there is a large overlap in TCR sequences between unrelated individuals, providing a pool of public antigen receptors between individuals that are enriched in response to a shared pathogenic exposure.

Thus, the present invention provides in one aspect a method for identifying public antigen receptors as biomarkers, based on the discovery that specific amino acid sequences of TCR or IG loci are recurrently produced in multiple individuals, and may be used in responding to the same pathogen or environmental exposure. The rearranged VDJ regions from millions of rearranged TCR β genes in the naïve and memory CD8 compartments were sequenced and the unexpected result was a high overlap between individuals, where more than half of all TCR β clonotype sequences in each individual were derived from a small but specific subset of all possible TCR β clonotype sequences. Thus, the present invention provides in one aspect a method for identifying public antigen receptors as biomarkers, based on the discovery that specific amino acid sequences of TCR or IG loci are recurrently produced in multiple individuals, and may be used in responding to the same pathogen or environmental exposure. The rearranged VDJ regions from millions of rearranged TCR β genes in the naïve and memory CD8 compartments were sequenced and the unexpected result was a high overlap between individuals, where more than half of all TCR β clonotype sequences in each individual were derived from a small but specific subset of all possible TCR β clonotype sequences. The proportion of public polypeptide sequences specific to an immune response carried in the memory compartment or the overall repertoire of an individual is defined as the "frequency" of public antigen receptor sequences for a specific immune response, within that individual. Comparison of shared sequences across individuals with common exposures allows determines

panels of public antigen receptors specific for a given exposure disease. The joint frequency of sequences associated with a specific immune response in the overall repertoire of unexposed individuals is adequately small allowing discrimination between exposed and unexposed individuals. These individuals could also be individuals who have broken tolerance to self in autoimmune disease, as compared to individuals who have not yet initiated the immune response to self. In a given tissue sample, frequency can be measured as the number or proportion of total sequences, the number or proportion of unique sequences, or even the identity of specific sequences, and can be compared to each sequence observed in individuals experiencing a given immune response. The tissue sample might be whole blood or subsets thereof, biopsy from a specific tissue or lesion, or specific cell types extracted from a biopsy specimen. Sequences might be derived from the tissue sample on the basis of either DNA or RNA.

In another aspect, the present invention provides biomarkers generated from exposure-specific clonotype sequences that are identified as a result of using the methods of the present invention. The exposure-specific clonotype sequences are detected from immune cells taken from a group of subjects with a disease. For the purposes herein, a group of subjects is two or more people and is also referred to as a "population". The immune cells can originate from a tissue where a mixture of immune cells is present and T cells or B cells are isolated by any suitable means. Polynucleotide sequences, such as DNA or RNA are isolated and prepared in manner suitable for amplification using oligonucleotide primers in a nucleotide amplification reaction. The clonotype sequences that are shared in common between individuals sharing either exposure or autoimmune disease ("cases") are identified. The process is repeated with a control population identified as either unexposed to a given pathogen, not experiencing autoimmune disease, or even a cohort of individuals with unknown exposure status in the case of rare exposures, to establish which sequences shared among the cases are specific to the exposure/condition. The biomarker is defined as the panel of specific antigen receptor sequences that are significantly enriched in the cases as compared to controls.

In another aspect, the present invention provides a method for using these biomarkers or exposure-specific clonotype sequences as diagnostics for disease or risk of disease in normal subjects, to detect a disease or risk of a disease. For example, a panel of exposure-specific clonotype sequences could be used to monitor either the initiation or the progression of an autoimmune disease.

Clonotype sequences may refer to either amino acid or nucleotide sequences. Clonotype amino acid sequences can be encoded by multiple possible clonotype nucleotide sequences e.g.,

degenerate nucleotide sequences, so the diagnostic criteria include the amino acid sequence and all possible nucleotide sequences giving rise to that amino acid sequence.

In certain embodiments of the present invention the method provides for collecting samples of immune cells from the subject. In other embodiments, the samples are collected and prepared independently and are not necessary for practicing the invention. In either case, one way of collecting samples uses whole blood from an individual where Peripheral Blood Mononuclear Cells (PBMCs) are isolated, for example, on a Ficoll gradient. The cells can be cryopreserved for later use, and screened for tetramer positive T cells for a specific antigen. Whole blood is processed to isolate memory B cells (CD19+, CD27+), memory CD4 T cells (CD4+, CD45RO+) and memory CD8 T cells (CD8+ CD45RO+). Apparatus for separation are known, for example the AutoMACS® magnetic bead cell separator (Miltenyi Biotec, Auburn, CA).

B cell or T cell nucleotides are sequenced. If total genomic DNA is to be used, it 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. Using PBMCs as a source, the number of T cells can be estimated to be about 30% of total cells. Alternatively, total nucleic acid can be isolated from cells, including both genomic DNA and mRNA. Methods for extracting nucleotides from immune cells and tissue are well known to those of ordinary skill in the art.

A variety of methods are useful to detect diagnostic clonotype sequences of either TCR or IG for a given disease state. Methods include the comprehensive sequencing strategy disclosed herein and used to discover exposure-specific clonotype sequences, but are not limited to this method as other sequence-specific detection methods available should work, including hybridization, TaqMan, and other technologies that detect the presence or absence of specific nucleotide sequences. Additionally, the peptides encoded by the diagnostic TCR or IG might be detected directly by immunoassay with either the appropriate antigen or monoclonal antibodies developed for this purpose. The immunoassays which can be used include, but are not limited to, competitive assay systems using techniques such western blots, radioimmunoassays, ELISA, "sandwich" immunoassays, immunoprecipitation assays, precipitin assays, gel diffusion precipitin assays, immunoradiometric assays, fluorescent immunoassays, protein A immunoassays, and complement-fixation assays. Such assays are routine and well known in the art (see, e.g., Ausubel *et al.*, eds, 1994 *Current Protocols in Molecular Biology*, Vol. 1, John Wiley & sons, Inc., New York). Additionally, routine cross-blocking assays such as those described in *Antibodies, A Laboratory Manual*, Cold Spring Harbor Laboratory, Ed Harlow and

David Lane, 1988), can be performed. In certain embodiments, nucleic acid based methods are preferred over protein based assays, but either can distinguish individuals carrying the clonotype-specific sequence at high frequency from individuals not carrying the diagnostic clonotype.

The present inventors previously disclosed a computational method to measure TCR CDR3 diversity based on single molecule DNA sequencing in WO2010/151416, which is hereby incorporated by reference in its entirety. In those studies, a method was developed 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 TCR β CDR3 sequences in the adult repertoire significantly exceeds previous estimates based on exhaustive capillary sequencing of small segments of the repertoire. The TCR β chain diversity in the CD45RO⁻ population (enriched for naïve T cells) observed using the methods described herein was five-fold larger than previously reported. The number of unique TCR β CDR3 sequences expressed in antigen-experienced CD45RO⁺ T cells was between 10 and 20 times larger than expected. The frequency distribution of CDR3 sequences in CD45RO⁺ cells suggested that the T cell repertoire contains a large number of clones with a small clone size. Furthermore, it was determined that the realized set of TCR β chains were 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) were created at a relatively high frequency. TCR sequences close to germ line were 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.

Additionally, the insertion rates of nucleotides were strongly biased. The frequency distribution of CDR3 sequences in CD45RO⁺ cells was notable for a significant number of sequences that were observed with low copy counts, suggesting that the $\alpha\beta$ CD45RO⁺ T cell repertoire contains a large number of clones with a small clone size.

Previous work demonstrated that the number of unique CDR3 sequences found in the repertoire of a single individual is $3-4 \times 10^6$ (WO2010/151416 and Robins, *et al.*, *Blood* 114, 4099-4107 (2009)), which, although large, comprises a negligible fraction of the number of the estimated 5×10^{11} theoretically possible TCR β CDR3 sequences. Based on this, one would expect fewer than 5 sequences shared between any two individual repertoires assuming $3-4 \times 10^6$ sequences present in each individual were randomly chosen from a uniform distribution of 5×10^{11} sequences. The present inventors compared the TCR β CDR3 sequence repertoires present in pairs of healthy adults and unexpectedly found that the actual repertoire overlap between any

two individuals is several thousand-fold larger than predicted. Moreover, the degree of overlap did not appear to depend strongly on the extent of HLA-A-, -B, or -C matching. The finding implied either (1) that individual TCR β sequence repertoires were not randomly selected from the space of possible sequences, or (2) that the sequences in this space were not uniformly probable, or both. Distinguishing between these two possibilities was important, because the first possibility would be consistent with convergent evolution during T cell development and the second implies that the sequence space of receptors in the cellular adaptive immune system is much smaller than presently believed. Convergent evolution is the possibility that a diverse set of TCRs rearrange in the thymus, and the positive and negative selection process favors the same lower diversity subset of TCRs in each individual. Arranging the TCR β CDR3 sequences observed in each donor according to the number of junctional insertions demonstrated that the sequences were not randomly sampled from the set of possible sequences. This led to the conclusion that majority of sequences in each donor sampled were from a small corner of the set of possible sequences, and that the distribution of sequences in the possible sequence space is not uniform. To determine whether the distribution was sufficiently non-uniform to account for the larger than expected observed overlaps between individual repertoires, a model that assumes that the distribution is uniform within each subset of sequences carrying a specified number of insertions was created. The model let the fraction of each subset (with a specified number of insertions) equal the empirically determined value. Using the model, the overlap expected when the CDR3 sequences are randomly sampled from this distribution was predicted. It was determined the computed overlap was ~14,000, which was indistinguishable from the observed overlap in the cohort. Thus, one does not need to invoke non-random selection of the sequences in the possible sequence space to account for the observed overlap. In other words, there was no evidence of convergent evolution.

The results determined that the small effective size of the TCR repertoire was primarily a result of the stochastic process of insertion of non-templated nucleotides. The number of nucleotides inserted into a given junction varied according to a probability distribution, which was empirically determined from the data. The sequences with low numbers of insertion were selected from a small number of possibilities and, were likely to be created in multiple people. The process of β -chain rearrangement occurs during T cell development, before the functional $\alpha\beta$ receptor was displayed on the surface of the T cell, prior to thymic selection.

The large shared overlap of TCR sequences suggested that multiple people were able to respond to the same antigen with identical (or very similar) T cell clones. These public

responses are believed to be common, as supported by the EBV example. The clinical application is the diagnostic utility of these overlapping TCR sequences. With the technology presented, one can readily detect the repertoire of expanded clones in the memory compartment. A set of "public" TCR clones induced by a particular antigen can be used to diagnose a disease state associated with the antigen. The HLA variability is a consideration in the application of such a diagnostic, but many diseases, such as Type 1 Diabetes are strongly associated with particular HLA alleles.

The biomarkers and methods of the present invention allow one of skill in the art to identify, diagnose, or otherwise assess subjects who do not exhibit any symptoms of exposure to a pathogen, or initiation of an autoimmune disease. For example, a subject may not have clinically symptomatic diabetes, in particular Type I diabetes, but nonetheless may be at risk for developing diabetes or experiencing symptoms characteristic of a diabetic condition. Other autoimmune diseases include, but are not limited to: multiple sclerosis, rheumatoid arthritis, scleroderma, Primary CNS Vasculitis, Rasmussen's Encephalitis, Autoimmune Peripheral Neuropathy, Autoimmune Cerebellar Degeneration, Gait Ataxia with Late Age Onset Polyneuropathy (GALOP), Stiff Person Syndrome, Chronic Inflammatory Demyelinating Polyneuropathy, Myasthenia Gravis, Lambert Eaton Myasthenic Syndrome, HTLV-1-Associated Myelopathy, (HAM) / Tropical Spastic Paraparesis (TSP), Opsoclonus / Myoclonus (Anti-Ri), and Neuromyelitis Optica, Grave's disease, Hashimoto's thyroiditis, Celiac disease, Crohn's disease, Ulcerative colitis, Systemic Lupus Erythematosus, Goodpasture's syndrome, Wegener's granulomatosis, Polymyalgia Rheumatica, Guillain-Barre syndrome, Addison's disease, Ankylosing Spondylitis, Psoriasis.

Infectious diseases that might generate useful biomarkers for exposure status include bacterial infections such as: *Escherichia coli*, *Haemophilus influenzae*, *Actinomyces*, *Listeria*, *Meningococcus*, *Pneumococcus*, *Pseudomonas*, *Salmonella*, *Shigella*, *Clostridium*, *Staphylococcus*, *Streptococcus*; viral infections including: Hepatitis A, Hepatitis B, Hepatitis C, Hantavirus, Dengue Fever, Herpes Simplex Virus, Herpes Zoster, Cytomegalovirus (CMV), Epstein-Barr virus (EBV), Ebola virus, Marburg Virus, SARS; fungal infections including: *Aspergillosis*, *Blastomycosis*, *Candidiasis*, *Coccidioidomycosis*, *Cryptococcosis*, *Histoplasmosis*, *Mucormycosis*, *Paracoccidioidomycosis*, *Sporotrichosis*; and parasitic infections including: *Amebiasis*, *Amebic Infections*, *Ascariasis*, *Babesiosis*, *Cryptosporidiosis*, *Dracunculiasis*, *Giardiasis*, *Hookworm Infection*, *Leishmaniasis*, *Malaria*,

Microsporidiosis, Onchocerciasis, Pinworm Infection, Schistosomiasis, Tapeworm Infection, Toxocariasis, Toxoplasmosis, Trichinosis, Whipworm Infection.

In addition to autoimmune disease and infectious disease, the present claims embody methods wherein other non-specific inflammatory diseases or immune dysfunction syndromes can include useful disease-specific antigen receptor sequences, such as: Gouty arthritis, Polymyalgia rheumatica, Kawasaki disease, juvenile dermatomyositis, DiGeorge Syndrome, severe combined immunodeficiency (SCID), common variable immunodeficiency (CVID), Bruton's agammaglobulinemia, AIDs caused by retroviruses HIV-1 or HIV-2. These lists of diseases and antigenic agents are illustrative, and in some cases preferred, but are not intended to be limiting.

Identifying a subject exposed to a pathogen, particularly a pathogenic antigen enables the selection and initiation of various therapeutic interventions or treatment regimens in order to delay, reduce or prevent that subject's onset of a disease state.

The present invention can also be used to screen subject populations in any number of settings. For example, a health maintenance organization, public health entity or school health program can screen a group of subjects to identify those requiring interventions, as described above, or for the collection of epidemiological data.

Tests to measure biomarkers and biomarker panels can be implemented on a wide variety of diagnostic test systems. Amplification of a selected a nucleic acid sequence may be carried out by a number of suitable methods. See generally Kwoh *et al.*, *Am. Biotechnol. Lab.* 8:14 (1990). Numerous amplification techniques have been described and can be readily adapted to suit particular needs of a person of ordinary skill. Non-limiting examples of amplification techniques include polymerase chain reaction (PCR), ligase chain reaction (LCR), strand displacement amplification (SDA), transcription-based amplification, the q3 replicase system and NASBA, all known to those skilled in the art.

Methods for performing immunoassays are well-established. Useful assays include, for example, an enzyme immune assay (EIA) such as enzyme-linked immunosorbent assay (ELISA), a radioimmune assay (RIA), a Western blot assay, or a slot blot assay (see, e.g., Diamandis, *Immunoassay* (Academic Press, Inc. 1996), U.S. Pat. Nos. 4,366,241; 4,376,110; 4,517,288; and 4,837,168).

The following examples describe the methods and compositions disclosed herein. In one example the subjects were positive for Epstein Barr Virus. In other examples, four hundred four (404) biomarkers were identified as being found be "public" across subjects who have Multiple

Sclerosis (MS) or exhibit symptoms characteristic of MS. Six hundred eighty-nine (689) biomarkers were also identified as being found “public” in subjects who have Type 1 diabetes, or who exhibit symptoms characteristic of diabetes, including, for example, insulin resistance or altered beta cell function.

The invention is illustrated by the following non-limiting examples.

EXAMPLES

I. Two donor study

Example 1: Sample acquisition, PBMC isolation, FACS sorting and genomic DNA extraction

The T-lymphocytes were flow sorted into four compartments for each subject: $CD8^+CD45RO^{+/-}$ and $CD4^+CD45RO^{+/-}$. The characterization of lymphocytes used the following conjugated anti-human antibodies: 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 FACS Aria™ cell-sorting system (BD Biosciences). Data were analyzed with FlowJo software (Treestar Inc.).

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 from each compartment was isolated, as shown in Table 1.

Table 1

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

Example 2: T cell receptor β chain spectratyping

TCR β chain spectratyping was performed as by those skilled in the art. 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 V β 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 μ M dNTPs, 2mM MgCl₂ and 1 unit of AmpliTaq Gold DNA polymerase (Perkin Elmer) in AmpliTaq Gold buffer, in a final volume of 20 μ l. After completion, an aliquot of the PCR product was diluted 1:50 and analyzed via capillary electrophoresis using a 3730X1 DNA Analyzer (Applied Biosystems). The output of the DNA Analyzer is a distribution of fluorescence intensity vs. time, which 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.

Example 3: Multiplex PCR amplification of TCR β CDR3 regions

In order to generate the template library for the Genome Analyzer, a multiplex PCR system to amplify rearranged TCR β loci from genomic DNA was designed. The multiplex PCR system uses 45 forward primers (SEQ ID. NOS:1-45), each specific to a functional TCR V β segment, and 13 reverse primers (SEQ ID. NOS:46-57), each specific to a TCR J β segment.

Primer design: 45 forward PCR primers (SEQ ID NOS: 1-45) complementary to each of the 48 functional Variable segments, and 13 reverse PCR primers (SEQ ID NOS: 46-57) complementary to each of the 13 functional Joining genes (TRBJ) from the TRB locus, as listed in the international ImMunoGeneTics information system® (See, Lefranc, M.-P. *et al.*, *Nucleic Acids Research*, 27, 209-212 (1999); Ruiz, M. *et al.*, *Nucleic Acids Research*, 28, 219-221 (2000); Lefranc, M.-P., *Nucleic Acids Research*, 29, 207-209 (2001); Lefranc, M.-P., *Nucleic Acids Res.*, 31, 307-310 (2003); Lefranc, M.-P. *et al.*, *In Silico Biol.*, 5, 0006 (2004); Lefranc,

M.-P. *et al.*, *Nucleic Acids Res.*, 33, D593-D597 (2005); Lefranc, M.-P. *et al.*, *Nucleic Acids Research* 37: D1006-D1012; 2009). 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).

The forward primers are modified at the 5' end with the universal forward primer sequence compatible with the GA2 cluster station solid-phase PCR. Similarly, all of the reverse primers are modified with the GA2 universal reverse primer sequence.

Analysis of existing data in the IMGT database shows that the average J deletion is 4 bp +/- 2.5 bp, which implies that J deletions greater than 10 nucleotides occur in less than 1% of sequences. The 13 different TCR J β gene segments each have a unique four base tag at positions +11 through +14 downstream of the RSS site. Thus, sequencing oligos are 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 (SEQ ID NOS: 58-70).

The information used to assign the J and V segment of a sequence read was entirely contained within the amplified sequence, and did not rely upon the identity of the PCR primers. The sequencing oligonucleotides 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 oligo. In this way, promiscuous annealing of the sequencing oligonucleotides should not impact the quality of the sequence data generated.

The average length of the CDR3 region, defined following convention as the nucleotides between the 2nd conserved cysteine of the V segment and the conserved phenylalanine of the J segment, is 35 +/- 3, so sequences starting from our J β segment tag nearly always captures the complete VNDNJ junction in a 50 bp read.

TCR β J gene segments are roughly 50 bp in length. PCR primers that anneal and extend to mismatched sequences are referred to as promiscuous primers. The design addressed the risk of promiscuous priming in the context of multiplex PCR, especially in the context of a gene family, the TCR J β Reverse PCR primers to minimize overlap with the sequencing oligonucleotides. Thus, the 13 TCR J β reverse primers are anchored at the 3' end on the consensus splice site motif, with minimal overlap of the sequencing primers. The TCR J β primers were designed for a consistent annealing temperature (58 degrees in 50 mM salt) using

the OligoCalc program under default parameters (Kibbe, *Nucleic Acid Res.* 35(Sup.2)43-46 (2007)).

The 45 TCR V β forward primers were designed to anneal to the V β segments in a region of relatively strong sequence conservation between V β 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 V β gene segment used. This obviates the risk of erroneous TCR V β gene segment assignment, in the event of promiscuous priming by the TCR V β primers. TCR V β forward primers were designed for all known non-pseudogenes in the TCR β locus.

The total PCR product for a successfully rearranged TCR β CDR3 region using this system was expected to be approximately 200 bp long. Genomic templates were PCR amplified using an equimolar pool of the 45 TCR V β F primers (the "VF pool") and an equimolar pool of the 13 TCR J β R primers (the "JR pool"). 50 μ l PCR reactions were set up at 1.0 μ M VF pool (22 nM for each unique TCR V β F primer), 1.0 μ M 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/ μ l 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 TCR β CDR3 loci.

Example 4: Estimating relative CDR3 sequence abundance in PCR pools and blood samples

After collapsing the data, the underlying distribution of T-cell sequences in the blood was reconstructed because of the sequence data output from the GA experiments. The experimental procedure used three steps; 1) flow sorting T-cells drawn from peripheral blood, 2) PCR amplification, and 3) sequencing. The observed data was worked from backwards 3) to first determine the ratio of sequences in the PCR distribution 2) before estimating the true distribution of clonotype sequences in blood.

For each sequence observed a given number of times in our data, estimates of the probability that that sequence was sampled from a particular size PCR pool were made. 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 blood.

Example 5: Unseen species model for estimation of true diversity

Limited sample volumes from donors required a mixture model to reconstruct the frequency of each TCR β CDR3 species drawn from the blood, particularly for patients undergoing treatment.

The mathematical solution devised provided that for a total number of TCR β “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 . If the number of T cell genomes sequenced are experiment 1, then the number sequenced in a second experiment is t where t is exact same experiment).

Since there are a large number of unique sequences, an integral was used instead of a sum. If $G(\lambda)$ is the empirical distribution function of the parameters $\lambda_1, \dots, \lambda_S$, and n_x is the number of clonotypes sequenced exactly x times, then

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

To find $\Delta(t)$, which is the number of *new* clonotype sequences observed in the second sequencing experiment.

$$\Delta(t) = \sum_x E(n_x)_{\text{exp1+exp2}} - \sum_x E(n_x)_{\text{exp1}} = S \int_0^{\infty} e^{-\lambda} (1 - e^{-\lambda t}) dG(\lambda)$$

Taylor expansion of $1 - e^{-\lambda t}$ gives $\Delta(t) = E(x_1)t - E(x_2)t^2 + E(x_3)t^3 - \dots$, which can be approximated by replacing the expectations ($E(n_x)$) with the observed numbers in the first experiment. Substituting in the numbers observed in experiment 1, this formula predicted that $1.6 \cdot 10^5$ new unique sequences should be observed in experiment 2. The actual observation was $1.8 \cdot 10^5$ new TCR β sequences, a valid lower bound on total diversity. The expression for $\Delta(t)$

oscillates wildly as t goes to infinity, however, so to produce a lower bound for $\Delta(\infty)$, $\Delta(t)$ needs to be regularized. There are many known methods to accomplish this, and e.g. Euler's transformation.

Example 6: $J\beta$ gene segment usage

The CDR3 region in each TCR β chain included sequences derived from one of the 13 $J\beta$ 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 13 different $J\beta$ gene segments varied more than 20-fold. The $J\beta$ gene segment usage pattern observed in the four different T cell populations was relatively constant within a given donor. Moreover, the $J\beta$ 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.

Example 7: TCR β chains with identical amino acid sequences found in different people

The TCR β chain sequences were translated to amino acids and then compared pairwise between our two donors. Many thousands of exact sequence matches were found. For example, comparing the $CD4^+$ $CD45RO^-$ sub-compartments, approximately 8000 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, 1500/8000 identical amino acid matches had >5 nucleotide mismatches. Between any two T cell sub-types, we find 4-5% of the unique TCR β sequences have identical amino acid matches.

To confirm our finding of thousands of identical TCR β chain amino acid sequences, we compared our two donors to the $CD8^+$ $CD62L^+$ $CD45RA^+$ (naïve-like) TCRs from a third donor who was also Caucasian, but a 44 year old CMV $^+$ female. Identical pairwise matches of many thousands of sequences at the amino acid level between our third donor and each of our original two donors were found (See Figure 1).

Example 8: Larger clonotypes are closer to germline

The variation in copy number between different sequences within every T cell sub-compartment that we studied ranged by over a factor of 10,000. The only property identified that correlated with copy number was the number of insertions + the number of deletions (with an inverse correlation).

Example 9: Estimation of total CDR3 sequence diversity

The number of unique CDR3 sequences observed in each lane of the GA 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 TCR β 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”, an analytic solution to which was developed over 60 years ago. The solution starts with determining the number of new species, or TCR β CDR3 sequences, that are observed if the experiment is repeated, i.e., if the GA sequencing is repeated on an identical sample of peripheral blood T cells. For our purposes, this amounts to sequencing an identically prepared library of TCR β CDR3 PCR products in a different lane of the GA 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%, suggesting that this analytic solution can, in fact, be used to estimate the total number of unique TCR β CDR3 sequences in the entire repertoire.

The resulting estimates of the total number of unique TCR β CDR3 sequences in the four flow cytometrically-defined T cell compartments are shown in Table 2.

Table 2: TCR repertoire diversity

Donor	CD8	CD4	CD45RO	Diversity	GA*
1	+	-	+	6.3×10^5	2
1	+	-	-	1.24×10^6	2
1	-	+	+	8.2×10^5	2
1	-	+	-	1.28×10^6	2
2	+	-	+	4.4×10^5	1
2	+	-	-	9.7×10^5	1
2	-	+	+	8.7×10^5	2
2	-	+	-	1.03×10^6	2

Of note, the total TCR β diversity in these populations is between 4-5 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 TCR β CDR3 repertoire sizes of each compartment in the two donors are within 20% of each other.

II. Seven Donor Study

Example 10: Sample acquisition, PBMC isolation, FACS sorting and genomic DNA extraction

A 100 blood samples of 200 ml whole blood from each individual are collected and antibody titers are measured at several laboratories. DNA is sequenced and amplified as described in Example 1.

The characteristics and demographic information of the donors was identified, and included: HLA-A, -B, -C typing data, and Epstein-Barr virus (EBV) serostatus

“+” indicates detectable serum IgG against the viral capsid antigen of EBV. Donors 1 and 3 were full siblings and the daughters of donor 2; the other four donors were unrelated.

Table 3: Donor characteristics

Donor	Age	Sex	Geographic Origin	HLA-A	HLA-B	HLA-C	EBV serostatus
1	25	F	Caucasian	*0201/*0101	*44/*57	*06/*16	+
2	51	F	Caucasian	*01/*24	*38/*57	*06/*12	+
3	22	F	Caucasian	*0201/*0101	*44/*57	*06/*16	+
4	39	M	Japanese	*26/*31	*07/*56	*01/*07	+
5	38	M	Indian	*0205/*24	*35/*35	*04/*12	+
6	20	F	African	*01/*34	*0801/*53	*04/*07	+
7	53	M	African	*23/*33	*57/*58	*10/*18	+

The serum tubes are processed according methods standard and known in the art for each of the serology assays specified in Table 3 above. PBMCs are isolated from the 50 ml whole blood aliquots on a Ficoll gradient. Cryopreserved PBMCs are prepared and can be screened for tetramer positive T cells for a specific antigen. Buffy coat DNA is isolated from the second syringe, for use in screening for specific sequences in the naïve cells. The third syringe of whole blood is processed on the AutoMACS magnetic bead cell separator to isolate approximately 10⁶ memory B cells (CD19⁺, CD27⁺), 10⁶ memory CD4 T cells (CD4⁺, CD45RO⁺) and 10⁶

memory CD8 T cells (CD8+ CD45RO+). DNA is prepared from each of the three cell subpopulations: B cell and CD4 T cell DNAs is sequenced.

HLA typing: Because CD4 sequences are included in the adaptive profile, high resolution typing at the HLA-DRB1 gene, to identify carriers of high frequency alleles at this locus, which may constrain the CD4 memory sequences is performed.

To profile past and ongoing exposures, not only acute exposures, only titer information for each pathogen in the IgG class of antibodies is analyzed. The IgM class is typically associated with very recent exposures. In order to maximize the statistical power for associating exposure status with adaptive immune profile, a panel of pathogenic exposures of between 10% and 90% seroprevalent is expected. These are detailed in Table 11. The panel encompasses 13 pathogens, and 40 serological subtypes.

Table 4: pathogen serology

Pathogen	Seroprevalence*	Comment
CMV	55%	
H pylori	25%	
HSV1	60%	HSV panel yields serology on two serotypes simultaneously
HSV2	35%	
HHV6	60-100%	
Coxsackie B	15-90%	6 serotypes
Parvovirus B19	50%	
Hepatitis A	20%	Hepatitis panel yields data on both A and B
Hepatitis B	5%	
C. pneumonia	50%	Chlamydia panel yields simultaneous data on all three species
C. trachomatis	55%	
C. psittaci	<5%	
S. pneumonia	>10%	Seroprevalence of 23 serotypes, some natural exposure but significant seroprevalence through vaccines

* Seroprevalence data (Mandell, Douglas, Bennett, 7th edition 9) and Taylor and Blaser, *Epidemiol Rev* 13, 42 (1991).

Example 11: Explicit generation of TCR β CDR3 sequences

The potential β chain repertoire in humans has been calculated as 10^{11} possible amino acid sequences. Much of the predicted TCR β diversity derives from non-templated nucleotide insertions at the V-D and D-J junctions. The prediction of six insertions at each of the two junctions, for a total of 12 non-templated nucleotides. The cumulative distribution of TCR

sequences is plotted as a function of number of insertions for the naïve and memory compartments, respectively, from seven donors. Although examples of beta chains with higher numbers of insertions were observed, these were rare; less than 5% of human TCR β chain sequences had 12 or more insertions. Surprisingly, over 10% of all observed TCR β chain sequences had either zero or one inserted nucleotide, and almost half of the TCR β sequences in each of our donors had five or fewer total insertions, including both junctions. This shows that a large fraction of the TCR β repertoire is sampled from a small fraction of the 10^{11} possible sequences.

To determine the full set of TCR β sequences with a fixed maximum number of insertions, the explicit generation the sequences for models with different numbers of insertions was performed, allowing for deletions of zero to ten nucleotides at each of the recombination signal sequences. Upon comparing these generated sequence lists to the real sequence data, it was found over half of all TCR β sequences in the population are contained on the list of sequences generated from a model with five or fewer insertions. The total number of sequences on this list is $\sim 2 \times 10^8$, so it appears that more than half of the TCR β repertoire is sampled from less than 0.2% of the 10^{11} possible sequences in a model allowing 12 or fewer insertions. The small effective size of the potential TCR repertoire implies that the overlap of the cellular adaptive immune system between individuals is far greater than present dogma would suggest.

All possible TCR β CDR3 nucleotide sequences containing a total of $N = 0, 1, 2, 3, \dots, 7$ nucleotide insertions at the V β D β and D β J β junctions were systematically generated with the model of VDJ rearrangement. The model was governed by a set of rules that were empirically determined from the observed TCR β CDR3 sequence data from the seven donors. The model allowed all possible V-D-J combinations, deletion of up to 10 nucleotides from the 3' end of the V β segment and from the 5' end of the J β segment, deletion of nucleotides from the 5' and 3' ends, up to complete deletion, of the D β segment, and up to 7 total junctional insertions. The total length of the CDR3 sequence, defined as the interval from the codon for the conserved cysteine at the 3' end of the V β gene segment to the codon for the conserved phenylalanine in the 5' portion of the J β gene segment, was constrained such that it could encode from 9 to 23 amino acid residues. Generation of the CDR3 amino acid sequences containing a total of 7 junctional insertions was performed on a 50-node linux cluster with 8 CPUs per node, and required 24 hours to complete. Each CDR3 nucleotide sequence was translated into the corresponding amino acid sequence and then stored in a binary tree organized by alphabet. For

each new sequence not found in the tree, a new leaf was created with that sequence. For each sequence already found in the tree, the count for the corresponding leaf was incremented by one.

Example 12: Utilization of V-J Usage

To determine whether the CD8⁺ TCR β CDR3 repertoire in each individual is randomly sampled from the set of 5×10^{11} possible sequences, explicit enumeration of the complete set of CDR3 sequences predicted by a model of VDJ recombination that allowed up to ten nucleotide deletions from the ends of the V β , D β and J β gene segments adjacent to the RSS, followed by insertion of a total of up to seven non-templated nucleotides at the V β -D β and D β -J β junctions was done. The model allowed the total CDR3 length to range from 27 to 69 nucleotides, encoding 9 to 23 amino acids, consistent with the experimentally observed sequence data. Generation of all unique CDR3 amino acid sequences containing a total of 7 or fewer nucleotide insertions at the V β -D β and D β -J β junctions required approximately 10,000 cpu hours on a 2.3 Ghz processor (Robins *supra*). Comparison of the set of sequences observed in the naïve and memory CD8⁺ repertoires of each of the seven donors with the full set of sequences generated by the model and revealed that $51.1 \pm 3.5\%$ of all the sequences observed in the naïve CD8⁺ compartment of each donor are found in the subset of 5.7×10^8 predicted sequences containing six or fewer total insertions. Thus, the majority of TCR β CDR3 amino acid sequences are drawn from a defined subset of less than 0.11% of the estimated 5×10^{11} possible TCR β sequences.

Much of the predicted diversity in the TCR β CDR3 repertoire is generated by non-templated nucleotide insertions at the V β -D β and D β -J β junctions. The cumulative distribution of TCR β CDR3 sequences observed in the CD8⁺ naïve and memory compartments, respectively, of the seven donors as a function of the number of junctional insertions demonstrates that sequences with 12 or more insertions were observed, but constitute only 10% of the total. In contrast, more than 10% of the observed sequences had zero, one, or two insertions, and 50% of the sequences in each donor had six or fewer total insertions at the two junctions.

Example 13: Determining the Overlap

The pairwise overlap of the naïve CD8⁺ subsets of the seven donors predicted by our model of TCR β VDJ rearrangement is $1.44 \times 10^4 \pm 1.66 \times 10^3$ sequences which agrees closely with the overlap calculated between all 21 possible pairs of the seven individuals studied (Figure

5A). CDR3 sequences that were shared between individuals had fewer inserted nucleotides than the mean number observed across the entire repertoire.

Calculation of expected overlap between TCR β CDR3 amino acid sequence repertoires

(1) Expected overlap if all 5×10^{11} TCR β CDR3 amino acid sequences are equally likely:

The general calculation is in section (2), below. Here $n \approx 1 \times 10^6$ is the total number of unique TCR β CDR3 amino acid sequences in the naïve CD8 $^+$ compartment and f is $n/(5 \times 10^{11})$.

$$E[O] \approx n * f$$

Therefore, the expected overlap was approximately 2 sequences, substantially less than the observed >10,000.

(2) Estimate of expected overlap from the empirical distribution of sequences by number of insertions: The expected overlap O is the sum of the expected overlaps from the sequences with different numbers of insertions labeled by k .

$$E[O] = \sum_{k=0}^{15} E[O_k]$$

An explicit generation of the set of all TCR β sequences with a fixed number k of insertions (Table 1, Example 1) was called T_k . An empirical determination of the fraction f_k of the number of TCR β CDR3 sequences found in blood carrying k junctional insertions was made. Multiplying $f_k * n \equiv n_k$, where n is the total number of TCR β CDR3 sequences in the naïve CD8 $^+$ compartment as determined in (2), resulted in n_k , the number of unique TCR β CDR3 sequences in the blood with fixed number of insertions k .

The estimate of the number of sequences with k insertions that overlap between two individuals was equivalent to the problem of drawing n_k elements twice from a total distribution with T_k elements and determining the number of matches. The assumption was that each sequence with a fixed number of insertions is equally likely.) The expectation was determined from an approximately binomial distribution:

$$E[O_k] = \sum_{i=1}^{n_k} i \binom{n_k}{i} f_k^i (1 - f_k)^{n_k - i}$$

We calculate this using the generating function:

$$M(t) = e^{ti} \binom{n_k}{i} f_k^i (1 - f_k)^{n_k - i} = [f_k e^t + (1 - f_k)]^{n_k}$$

$$E[O_k] = M^{(0)} = n_k f_k$$

$$E[O] \approx \sum_{k=0}^{15} n_k f_k$$

Inserting the empirical data, it was determined that $E[O] \approx 1.54 * 10^4$, which is consistent with the observed pairwise overlap.

Example 14: Estimation of the actual overlap between different TCR β CDR3 repertoires

Previously established a lower bound for the total number of unique TCR β CDR3 amino acid sequences expressed in the naïve and memory CD8 $^+$ T cells circulating in the peripheral blood of a healthy adult (H. S. Robins, *et al.*, *Blood* 114, 4099-4107 (2009)). Only a fraction of these sequences can be observed experimentally, because the volume of blood that can be sampled at one time represents a small fraction of the total blood volume. The sets of TCR β CDR3 sequences that are observed experimentally in each donor are enriched for sequences expressed in CD8 $^+$ T cell clones that are common in the blood. Estimating the CDR3 sequence overlap between two entire T-cell repertoires must therefore take into account the shape of the empirically observed distribution to avoid over-estimation. To estimate the total number of TCR β CDR3 sequences that are shared between the CD8 $^+$ T cell compartments of two individuals, a non-linear regression approach was used in which the observed overlap data with a simple two parameter model, $Y = aX^b$, where x is the input variable described below, y is the number of overlapping sequences, and (a, b) are parameters to be estimated from the data.

First the unique TCR β CDR3 amino acid sequences encoded by the in-frame were sorted, read-through TCR β CDR3 nucleotide sequences observed in each donor in descending order according to their observed relative frequency. Let N_j be the total number of amino acid sequences in donor j , ($j = 1, 2$), and let n_{ij} be the top (or most frequent) n_i sequences from donor j , where i indexes the i th sequence. Then define the i th input variable, $X_i = n_{i1} * n_{i2}$, as the area equal to the maximum possible number of overlaps between n_{i1} and n_{i2} sequences. The outcome Y_i is the observed number of overlaps between in the n_{i1} sequences from donor 1 and the n_{i2} sequences from donor 2. The model parameters (a, b) are chosen to minimize the error between the observed number of overlaps and the regression model using the non-linear least squares method. Previously identified $\sim 1 \times 10^6$ as a lower bound for the number of unique

TCR β CDR3 amino acid sequences in the naïve CD8 $^{+}$ T cell compartment of a healthy adult was applied (Robins *supra*.) To estimate the minimum overlap between the entire naïve CD8 $^{+}$ repertoires of two donors, compute $Y = aX^b$ for $i = 1 \times 10^6$ and therefore $X_i = n_{i1} * n_{i2} = 1 \times 10^{12}$. To estimate the overlap between the naïve and memory CD8 $^{+}$ compartments of a single donor, however, use $i = 1 \times 10^6$ for the naïve compartment and $i = 3 \times 10^5$ for the memory compartment, since the lower bound for the number of unique TCR β CDR3 amino acid sequences in the memory CD8 $^{+}$ compartment is $\sim 3 \times 10^5$ (2). The calculated overlaps between the naïve and memory CD8 $^{+}$ TCR β CDR3 sequence repertoires for the full set of $7!/(2!5!)=21$ pairwise comparisons are summarized in Figure 1.

Example 15: Differential VJ Usage

The frequency with which specific V β -J β combinations were utilized was highly variable in each of the seven individuals. Although every possible V β -J β combination was observed, the frequency of specific combinations varied by more than 10,000-fold. V β -J β utilization was surprisingly consistent between individuals, especially for the rare V β -J β pairs, as reflected by the fact that the variance in V β -J β utilization was proportional to mean utilization. A fraction of the TCR β CDR3 sequences in the genomic DNA from the naïve and memory CD8 $^{+}$ T cells of each of the seven donors was predicted to generate out-of-frame TCR β transcripts that do not encode functional TCR β chains (Table 5). The V β -J β utilization in the out-of-frame CDR3 sequences was highly non-uniform and qualitatively similar to that observed for in-frame transcripts. The variability of V β -J β utilization in the out-of-frame CDR3 sequences cannot be attributed to positive or negative selection in the thymus of T cells bearing specific receptors, because these sequences do not generate proteins that participate in the selection process. The similarity in the utilization of specific V β -J β combinations in out-of-frame, nonfunctional and in-frame, functional TCR β transcripts therefore suggests that the variability in V β -J β utilization in both sets of sequences is attributable, at least in part, to mechanisms that operate before the stage of thymic selection.

The observed frequencies of specific V β -D β -J β combinations suggest that rearrangement between V β and D β gene segments is random, while that between D β and J β gene segments is not. The apparent non-random association between specific D β and J β gene segments is likely attributable to the organization of the TCR β locus, in which D β 1 lies 5' of all 13 J β segments,

while D β 2 lies 3' of the 6 members of the J β 1 cluster but 5' of the 7 members of the J β 2 cluster. The D β 1 segment is observed at roughly equal frequency with all 13 J β 's, while D β 2 is much more frequently paired with members of the J β 2 compared with the J β 1 family. D β 2 is observed with members of the J β 1 family about a third (.30+/-0.05) as often as would be expected if the pairing were random.

Table 5

Donor	Naïve Compartment	Naïve Compartment	Memory Compartment	Memory Compartment
	Unique Sequences	Total Reads	Unique Sequences	Total Reads
1	376,979	5,090,291	49,242	9,847,026
2	580,039	6,102,044	32,684	2,018,742
3	650,392	6,179,899	25,060	904,022
4	436,097	5,053,253	95,492	9,120,905
5	388,687	5,791,008	125,977	9,862,352
6	276,271	5,724,974	87,412	8,781,093
7	232,438	5,865,404	67,054	9,447,175

Example 16: Epstein Barr Virus exposure-specific clonotype sequences

A wide variety of pathogens are endemic to the human population, some of which are quite pervasive. A well studied example is the Epstein-Barr virus (EBV). A common, public TCR that responds to EBV in HLA B8 individuals has been described, with the V β CDR3 sequence CASSLGQAYEQYF, derived from V#, D# and J# with # inserted nucleotides (Argaet *et al.*, *J Exp Med* 180:2335-40, 1994, hereby incorporated by reference in its entirety). A2 specific clonotype sequences found with high frequency were:

CSARDRTGNGYT (SEQ ID. NO:71)

CSARDRVGNTIY (SEQ ID. NO:72)

CSARIGVGNTIY (SEQ ID. NO:73)

CSVGTGGTNEKLF (SEQ ID. NO:74)

A B8 specific sequence found with high frequency was CASSLGQAYEQYF (SEQ ID. NO:1167)

Of the seven donors, this TCR β clonotype was observed in the naïve CD⁺ compartment of two individuals, but in the memory compartment of only one donor, where it accounted for over 1% of all clonotype sequences. In the study a comparison of our donors' sequences with results from a study that identified 50 TCRs that interact with known EBV epitopes presented by HLA A-2 tetramers was performed. Four individuals matched at least one of these sequences in their naïve compartments. The two donors in our study that carried HLA A2 were the only individuals with matches to the public EBV responses in their memory compartment, and each of these two matched three of the 50 sequences. Thus, a total of four to seven HLA-restricted, public EBV responses were observed in our sample, and few had less than five inserted nucleotides.

Mice deficient for terminal deoxynucleotidyl transferase, the enzyme that catalyzes the template-independent insertion of nucleotides at the junctions, have 10-fold less diversity in their TCR CDR3 repertoires, with few insertions, yet these mice appear healthy, make efficient and specific immune responses, and display no increased susceptibility to infection (Gilfillan, *et al.*, *Eur J Immunol* 25, 3115-3122 (1995).; Cabaniols, *et al.*, *J Exp Med* 194, 1385-1390 (2001). 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, and that V β , D β , and J β segment sequences contributing to recurrently generated TCRs were be subject to evolutionary pressures favoring sequences recognizing antigens from common pathogens, as these sequences are present in the germline. Since germ line sequences are shared between people, it was hypothesized that shared TCR β chains are created by TCRs with a small number of insertions and deletions. Components of the CD8⁺ T cell response to ubiquitous pathogens such as Epstein Barr virus (EBV) are characterized by highly conserved TCR β CDR3 amino acid sequences that are found in multiple individuals and encoded by nucleotide sequences with few junctional insertions (Venturi, *et al.* *Nat Rev Immunol* 8, 231-238 (2008); Argae, *et al.*, *J Exp Med* 180, 2335-2340 (1994), Venturi, *et al.*, *J Immunol* 181, 7853-7862 (2008)).

Example 17: Correlations with Disease

To find a correlation with disease, investigation of 12 “public” TCR β CDR3 sequences that have been associated with the CD8⁺ response to EBV in individuals who express either HLA-A*0201 or HLA-B*0801 and detected 5 HLA-A*0201-associated, EBV-specific CDR3 sequences in the memory CD8⁺ compartments of donors 1 and/or 3, both of whom are HLA-

A*0201⁺, and an HLA-A*0801-associated, EBV-specific CDR3 sequence in the memory compartment of donor 7, who is HLA-B*0801⁺ was made. None of these responses were detected in the other four donors, all of whom were HLA-A*0201⁻ and HLA-B*0801⁻. The observation of the HLA-A*0201- and HLA-B*0801-associated, EBV-specific CDR3 sequences only in the three donors expressing one of the associated HLA alleles was statistically significant ($P = 0.0002$ by two-tailed Fisher exact test;).

Evaluation of correlation between observation of “public” EBV-associated TCR β CDR3 sequences and expression of the corresponding class I MHC allele.

A two-sided Fisher Exact Test was utilized to test the hypothesis that there is a correlation in our sequence data between the observation of 12 “public”, *i.e.*, found in multiple individuals, EBV-associated TCR β CDR3 amino acid sequences reported in the literature and expression of their associated class I MHC restricting elements. Previous studies have identified at least 11 public TCR β CDR3 amino acid sequences used by CD8⁺ T cells specific for EBV-encoded peptides presented by HLA-A*0201 (Venturi, *et al.*, *J Immunol* 181, 7853-7862 (2008)), and one public CDR3 sequence specific for an EBV-encoded peptide presented by HLA-B*0801 (Argaet *supra*). The seven individuals studied included two individuals expressing HLA-A*0201 and one individual expressing HLA-B*0801 (Table 1, Example 1). Thus, there were 2*11=22 CDR3 sequence:HLA-A*0201 combinations and one CDR3 sequence:HLA-B*0801 combination possible in the dataset that would support a correlation between the observation of a public EBV-specific CDR3 sequence and expression of the associated MHC restricting allele. Similarly, there were 5*11=55 possible CDR3 sequence:HLA-A*0201 combinations and 6*1=6 possible CDR3 sequence:HLA-B*0801 combinations, for a total of 61 combinations, that would *not* support a correlation. Five of the public EBV-specific CDR3 sequences associated with HLA-A*0201 in the memory CD8⁺ compartments of one or both of the HLA-A*0201-*positive* donors but none of the five HLA-A*0201-*negative* donors, and observed the public EBV-specific CDR3 sequence associated with HLA-B*0801 in the memory CD8⁺ compartment of the sole HLA-B*0801-*positive* donor but in none of the HLA-B*0801-*negative* donors were observed. Thus, 6 of the 23 possible public CDR3 sequence:MHC allele combinations that would support a correlation, and none of the 61 possible public CDR3 sequence:MHC allele combinations that would *not* support a correlation, were observed in the data. To test the null hypothesis of no correlation, the Fisher Exact Test was used which calculates a p-value of 0.0002, and provided confidence to reject the null hypothesis.

Table 6: False positive rates for sample size

number of cases	fraction with "public" BCR	N sharing "public" BCR	P(false positive)
30	25%	7.5	0.0119
50	25%	12.5	0.0172
70	25%	17.5	0.1946
30	50%	15	1.43E-06
50	50%	25	2.98E-06
70	50%	35	3.79E-04

Table 6 presents the probability of finding a false positive as a function of signal size. For the expected range of signal, a proposed study is sufficiently powered to limit false positives and to detect any sequence that is found in 20% or more the cases. Note that Table 6 is the marginal calculation for each disease. Taking all the diseases together, a multiple hypothesis test correction is needed. The Benjamini Hochberg FDR analysis is sufficient.

III. Additional Testing

Example 18: Diabetes

A study in Type 1 diabetic patients was done using protocols described above. Type 1 diabetes associates strongly with the class 2 HLA DRB1*03/04 genotype, so we sorted for CD4 memory (CD4+, CD45RO+) cells, and screened the CD4 memory cells for public sequences shared between three cases of T1D. These "public" TCR sequences were found in less than 5 of 10 HLA matched controls without T1D. DNA from all 13 donors was sequenced for both memory and naïve genotypes. The clonotype sequences that distinguish the T1D cases from the controls are shown as SEQ ID. NOS: 75-763.

Example 19: Multiple Sclerosis

Similar to the example above, CD4 memory cells were sorted from three MS cases carrying the DRB1*1501 allele, as well as three HLA-matched controls, to identify public sequences that might be enriched among MS cases. A set of "public" T cells were selected as TCR sequences found in all three cases and no more than one of the three controls. DNA from all donors was sequenced for both memory and naïve genotypes. The clonotype sequences that distinguish the MS cases from the controls are shown as SEQ ID. NOS: 764-1166.

What is Claimed:

1. A method for identifying a biomarker for a disease comprising:
 - a) providing isolated polynucleotide sequences from immune cells from a group of patients with the disease;
 - b) performing a nucleotide amplification reaction to produce a set of clonotype sequences;
 - c) identifying clonotype sequences enriched within the group of patients;
 - d) providing isolated polynucleotide sequences from immune cells from a group of normal subjects without the disease, and amplifying the polynucleotides;
 - e) removing sequences present in the normal subject group, which are obtained in step d), from the exposure-specific clonotype sequences, which are obtained in step c).
2. The method of claim 1, wherein the polynucleotide sequences are rearranged genomic sequences.
3. The method according to claim 1, wherein the nucleotide amplification reaction comprises:
 - (i) 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
 - (ii) 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 amplify a TCR CDR3 region.
4. The method according to claim 3, wherein the V- segments are selected from the group consisting of V α , V β , V γ and V δ , and the J-segments are selected from the group consisting of J α , J β , J γ and J δ .
5. The method of claim 3, wherein the immune cell samples are selected from the group consisting of CD45RO⁺, CD45RA^{int/neg} CD8⁺ T cells and CD45RO⁻ CD45RA^{hi} CD62L^{hi} CD8⁺ T cells.

6. The method of claim 5, wherein the CD8 cells are CD45RO⁺, CD45RA^{int/neg} CD8⁺ T cells.
7. The method of claim 5, wherein the T cells share one or more HLA alleles.
8. The method according to claim 1, wherein the nucleotide amplification reaction comprises:
 - (i) 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
 - (ii) a multiplicity of J primers, wherein each primer comprises a sequence that is complementary to a J-segment,wherein the V-segment and J-segment primers amplify an IGH , IGL or IGK CDR3 region.
9. The method of claim 3 or 8, wherein the V-segment primers (forward primers) are anchored at a position between 40 and 60 base pairs 5' of the recombination signal sequence (RSS), within the V segment.
10. The method of claim 3 or 8, wherein the J-segment primers (reverse primers) are at a position about 30 base pair 3' of the J gene RSS site.
11. The method of claim 1, wherein the exposure-specific clonotype sequences have an insertion of less than six nucleotides.
12. The method of claim 1, wherein the disease is selected from the group consisting of an autoimmune disease, an inflammatory disease, an immune deficiency, a bacterial infection, a viral infection, a fungal infection, and a parasitic infection.
13. The method of claim 1, further comprising an additional step prior to step (a) of isolating polynucleotide sequences from the immune cell samples.
14. The method of claim 13, wherein the polynucleotide sequences are isolated from an immune cell sample, wherein the sample is a tissue comprising hematopoietic lineage cells.

15. A biomarker produced by the method of claim 1.
16. A polypeptide encoded by the exposure-specific clonotype sequence of claim 15.
17. The polypeptide of claim of 16, wherein the polypeptide is antibody.
18. A diagnostic kit for detecting a disease or risk for a disease comprising one or more biomarkers produced by claim 1.
19. The diagnostic kit of claim 18, wherein the biomarker is selected from the group consisting of a polynucleotide biomarker, a labeled polypeptide biomarker, and an antibody biomarker.
20. The diagnostic kit of claim of 18, wherein the biomarker detects a disease or risk of a disease selected from the group consisting of an autoimmune disease, an inflammatory disease, an immune deficiency, a bacterial infection, a viral infection, a fungal infection, a parasite infection, and a prenatal disease.
21. A method for detecting a disease or a risk for a disease in a subject comprising:
 - a) selecting a biomarker for the disease according to the method of claim 1, wherein the biomarker consists of at least one exposure-specific clonotype sequence;
 - b) providing isolated polynucleotide sequences from immune cells of the subject, and amplifying the polynucleotides,
 - c) determining if the biomarker is present in the amplified polynucleotides of the subject.
22. The method of claim 21, further comprising determining that the presence of the biomarker in the amplified polynucleotides of the subject is indicative of the disease or the risk of the disease.
23. The method of claim 21, wherein the clonotype sequences are genomic sequences.

24. A method for detecting a disease or a risk of a disease in subject in a subject comprising:
- a) selecting a biomarker for the disease according to the method of claim 16;
 - b) detecting the presence of the biomarker in the polypeptide sequences of the subject.
25. The method of claim 24, further comprising determining that the presence of the biomarker in the polypeptide sequences of the subject is indicative of the disease or the risk of the disease.
26. A method for selecting a biomarker for diabetes or a risk for Type I diabetes comprising selecting a biomarker according to the method of claim 1, wherein the exposure-specific clonotype sequences are selected from the group consisting of SEQ ID. NOS: 75 to 763.
27. A method for detecting diabetes Type I or a risk for diabetes Type I in a subject comprising:
- a) selecting a disease-specific biomarker according to claim 26, wherein the biomarker consists of a disease specific clonotype sequence or a panel of such sequences;
 - b) providing isolated polynucleotide sequences from immune cells of the subject, and amplifying the polynucleotides,
 - c) determining if the of the diagnostic clonotype sequences are present in the amplified polynucleotides of the subject.
28. A method for detecting multiple sclerosis or a risk for multiple sclerosis in a subject comprising selecting a biomarker according to the method of claim 1, wherein the exposure-specific clonotype sequences are selected from the group consisting of SEQ ID. NOS: 764 to 1166.
29. A method for detecting multiple sclerosis or a risk for multiple sclerosis in a subject comprising:
- a) selecting a disease-specific biomarker according to claim 28, wherein the biomarker consists of a disease specific clonotype sequence or a panel of such sequences;

- b) providing isolated polynucleotide sequences from immune cells of the subject, and amplifying the polynucleotides,
- c) determining if the of the diagnostic clonotype sequences are present in the amplified polynucleotides of the subject.

1/1

Figure 1

