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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 February 2006 (09.02.2006)

PCT

(10) International Publication Number
WO 2006/014498 A2

(51) International Patent Classification:
C40B 40/10 (2006.01)

(21) International Application Number:
PCT/US2005/024002

(22) International Filing Date: 6 July 2005 (06.07.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/585,931 6 July 2004 (06.07.2004) US

(71) Applicant *(for all designated States except US)*: **BIOREN, INC.** [US/US]; 100 Glenn Way, Suite #1, San Carlos, CA 94070-6264 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **CREA, Roberto** [IT/US]; 700 Occidental Avenue, San Mateo, CA 94402 (US).

(74) Agents: **REMILLARD, Jane, E.** et al.; Lahive & Cockfield, LLP, 28 State Street, Boston, MA 02109 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2006/014498 A2

(54) Title: UNIVERSAL ANTIBODY LIBRARIES

(57) Abstract: Universal antibody libraries are described which are synthetic and derived from expressed human antibody sequences selected accordingly to certain criteria, for example, that the sequences are derived from naturally-occurring antibodies expressed in response to a certain antigen class (e.g., small molecule, polysaccharide, peptide, or protein) and having CDR regions engineered for optimal diversity. Methods for making and screening such libraries for isolating therapeutics suitable for treating disease are also disclosed.

UNIVERSAL ANTIBODY LIBRARIES

Related Information

The application claims priority to U.S. provisional patent application number 60/585,931, filed on July 6, 2004, the entire contents of which are hereby incorporated by reference. The contents of all patents, patent applications, and references cited throughout this specification are hereby incorporated herein by reference in their entireties.

Background of the Invention

Antibodies have profound relevance as research tools and in diagnostic and therapeutic applications. However, the identification of such useful antibodies is difficult and frequently, especially if therapeutic applications are envisioned, requires considerable redesign or 'humanization' before the antibody is suitable for administration.

Previous methods for identifying desirable antibodies have typically involved phage display of representative antibodies, for example human libraries or synthetic libraries, however, these approaches have limitations. For example, most human libraries contain only the antibody sequence diversity that can be experimentally captured or cloned from the source tissue. Accordingly, the human library may lack or under represent other valuable antibody sequences. Synthetic or consensus libraries have other limitations such as the potential to encode non-naturally occurring sequence that has the potential to be immunogenic. In addition, synthetic libraries, in an effort to be comprehensive, frequently contain too much diversity and are difficult to screen. Moreover, these libraries, when used to identify a candidate antibody that binds to a particular target, are not amenable to rational, follow-up, affinity maturation techniques to improve the binding of the candidate molecule. For example, methods for subsequent antibody improvement often involve *in vitro* mutagenesis such as random mutagenesis, saturation mutagenesis, error-prone PCR, gene shuffling, and antibody chain shuffling. These strategies are inherently stochastic and often require the construction of exceedingly large libraries to explore any meaningful sequence diversity. As the number of positions to be mutated in a given antibody becomes larger, the size of the resultant library becomes larger than what can be feasibly screened.

Accordingly, a need exists for a universal antibody library that systematically represents candidate antibodies that are non-immunogenic and have desired properties, for example, a representative diversity that can be readily screened.

Summary of the Invention

The invention solves the above problems by providing a universal antibody library (UAL) that represents all desirable candidate antibodies against a given antigen class as well as methods of making and screening such antibody libraries. Moreover, the antibodies of the universal antibody library are derived from human sequence, and are therefore nonimmunogenic, and therefore suitable for therapeutic applications, for example, for administering to human patients for preventing or treating human disorders or disease.

The libraries of the invention, have a diversity that is efficiently introduced using, for example, mutagenesis techniques such as walk-through mutagenesis (WTM) or look-through mutagenesis (LTM) (see respectively, *e.g.*, U.S. Pat Nos. 6,649,340; 5,830,650; 5,798,208; and U.S.S.N. 60/483,282) depending on whether multiple residue diversity or a single residue diversity needs to be introduced at a given site, *e.g.*, within one or more complementarity determining regions (CDRs). Importantly, these techniques allow for maximizing the amount of productive diversity and minimizing the amount of non-productive diversity, *i.e.*, mere noise or randomness. Accordingly, the universal antibody libraries of the invention can be smaller than existing antibody libraries yet comprise more rational diversity in order to identify candidate antibody binding molecules more efficiently.

In one embodiment, the universal antibody library of the invention is the application of the WTM or LTM technology to create a completely synthetic library that displays a desired diversity in one or more CDRs of the light and/or heavy chains. The antibody sequences, for example, the frameworks and CDRs are selected according to certain criteria. For example, one criterion is that the antibody sequence must have a minimum threshold frequency (*e.g.*, about 10% or more) of occurring within expressed (rearranged) antibody sequences, *e.g.*, human antibody sequences, and preferably, in response to a particular class of antigens. Optionally, yet another criterion, is that the expressed (rearranged) antibody sequence originates (or is derived from) with a minimum threshold frequency (*e.g.*, about 10% or more) from a germline sequence. Still another criterion/criteria is to make a comparison between CDRs (for example, expressed CDRs), that are of a given length, canonical structure, and/or CDR interdependency (e.g., CDR 1 against CDR2, and/or 3). The diversity is identified and then engineered into a conventional gene format, *e.g.*, a single chain antibody format (scFv), using oligonucleotides which allow for the complete assembly of framework and CDR sequences by genetic engineering (e.g., polymerase chain reaction (PCR), single overlap extension (SOE), and/or Kunkel-mediated mutagenesis), in a systematic manner.

Importantly, the invention minimizes any mutations that may lead to non-functional proteins by avoiding unwarranted mutations that typically occur when using mixed probes. In addition, the level of precision capable when using WTM contrasts with random mutagenesis and/or gene shuffling technologies. Moreover, by controlling framework selection and the level of sequence diversity in terms of position and amino acid type, the library's recognition of "antigen" classes is optimized. Furthermore, this *in vitro* methodology circumvents immunological negative selection of self-antigens and any gene bias due to the organism's environmental exposure.

Accordingly, the invention provides the advantage of being able to start with a screening library sized to be informative without being unnecessarily large. After the identification of the first set of clones, subsequent affinity maturation libraries can share common sets of LTM and/or WTM oligonucleotides saving time and reagent costs. Still further, the universal antibody libraries are capable of rapidly and effectively producing very specific antibodies against a variety of antigens, especially, *e.g.*, self-antigens which are difficult to obtain by any other method.

The universal antibody library is generated and screened by first synthesizing individual polynucleotides encoding a defined region or regions of an antibody where, collectively, the polynucleotides represent all possible variant antibodies according to the criteria described herein. The antibodies are expressed, for example, using *in vitro* transcription and translation and/or using a display technology, such as ribosome display, phage display, bacterial display, or yeast display.

The expressed antibodies are then screened and selected using functional assays, such as binding assays. In one embodiment, the polypeptides are expressed in association with the polynucleotide that encodes the antibody binding molecule, *e.g.*, a single chain antibody (scFv), thereby allowing for identification of the polynucleotide sequence that encodes the antibody binding molecule (*e.g.*, scFv). In a related embodiment, the antibodies are secreted and displayed on the membrane of a prokaryote such as *E. coli*, using, *e.g.*, the technology as described in, *e.g.*, US20040072740A1; US20030036092A1; and US20030100023A1.

The method can be used to identify human antibody sequences to develop new or improved antibodies or fragments thereof, *e.g.*, single chain antibodies (scFv). In addition, the method can be performed with the benefit of *a priori* information, *e.g.*, *via* computer modeling and electronic database biomining, that can be used to select an initial subset of sequences to be diversified, *e.g.*, according to the criterion described herein, using, *e.g.*, WTM or LTM mutagenesis.

Other advantages and aspects of the present invention will be readily apparent from the following description and examples.

Brief Description of the Figures

Figure 1 shows a schematic for carrying out the construction of a universal antibody library of the invention using computer-assisted database biomining.

Figure 2 highlights exemplary steps (and various database statistics) for identifying and selecting CDR and framework components for use in the synthesis of universal antibody libraries.

Figures 3-4 show exemplary threshold frequency of occurrence analyses, *i.e.*, an identification of the most often used germ line frameworks used in a human antibody immune response to a given antigen class. In ***Figure 4***, the relative frequency of germline contribution to rearranged antibodies is tabulated within each VH germline family.

Figure 5 shows the sequence of seven high frequency heavy chain frameworks used in response to a given antigen class (*e.g.*, a protein-based antigen) and their arrangement for functioning as acceptors for synthetic CDR regions. CDR are according to Contact definition (MacCallum *et al.*). Illustrated are enumerated germline V segments from VH1, VH3 and VH4.

Figure 6 shows the sequence diversity of an exemplary synthetic heavy chain CDR1 in the form of a CDR variability profile (frequency distribution). CDR1 length size 5 according to Kabat CDR definition.

Figure 7 shows the generated sequence diversity of an exemplary synthetic heavy chain CDR1 in the form of a CDR variability profile and a matrix showing residue positions and potential diversity. CDR1 length size 6 according to Contact CDR definition.

Figure 8 shows the sequence diversity of an exemplary synthetic heavy chain CDR2 in the form of a CDR variability profile (frequency distribution). CDR2 length size 17 according to Kabat CDR definition.

Figure 9 shows the generated sequence diversity of an exemplary synthetic heavy chain VH1 and VH3 CDR2 in the form of a CDR variability profile and a matrix showing residue positions and potential diversity. CDR2 length size 13 according to Contact CDR definition.

Figure 10 shows VH CDR3 length distribution of sizes 9 to 18 amino acids which cover about 75% of the available CDR space. A separated analysis was performed for each length (see Fig. 12). VH CDR3 sizes are according to Contact CDR definition

Figure 11 shows the sequence diversity of an exemplary synthetic heavy chain CDR3 in the form of a CDR variability profile (frequency distribution). CDR3 length size 13 according to Kabat CDR definition

Figure 12 shows the generated sequence diversity of an exemplary synthetic heavy chain CDR3 in the form of CDR variability profile and a matrix showing residue positions and potential diversity. CDR3 length sizes 9-18 according to Contact CDR definition.

Figure 13 shows the sequence diversity of each heavy chain CDR as well as the combined heavy chain library diversity. The number of variable positions and CDR sizes are according to Kabat definition.

Figure 14 shows the construction of the heavy chain library using a combination of overlapping nondegenerate and degenerate oligonucleotides which can be converted to double-stranded nucleic acids using the single overlap extension polymerase chain reaction (SOE-PCR).

Figure 15 shows the combining of the heavy chain library of the universal antibody library with a kappa and/or lambda light chain library for additional diversity.

Figure 16 shows the sequence of seven high frequency light chain frameworks (*i.e.*, three kappa and four lambda light chain frameworks) used in response to a given antigen class (*e.g.*, protein) and their arrangement for functioning as acceptors for synthetic CDR regions. CDR regions are identified according to Contact definition. Illustrated are enumerated germline V segments from Vkappa I-L1, Vkappa III-A27, Vkappa III-L6, Vlambda 1-1b, Vlambda 2-2a2, Vlambda 3-3l, and Vlambda 3-3r.

Figure 17 shows the generated sequence diversity of exemplary synthetic Vkappa I and Vkappa III light chain CDR1s in the form of variability profiles (frequency distributions) and permutation matrices. CDR1 length size 7 and 8 according to Contact CDR definition.

Figure 18 shows the nucleic acid and amino acid sequence of an exemplary single chain antibody (scFv) of the invention.

Figure 19 shows the generated sequence diversity of exemplary synthetic Vkappa I and Vkappa III light chain CDR2s in the form of variability profiles (frequency distributions) and matrix showing residue positions and potential diversity. CDR2 length size 10 according to Contact CDR definition.

Figure 20 shows Vkappa CDR3 length distribution of sizes 8 and 9 amino acids which cover about 80% of the available CDR space. A separated analysis was performed for each length (see Fig. 21). VH CDR3 sizes are according to Contact CDR definition.

Figure 21 shows the generated sequence diversity of exemplary synthetic light chain (Vkappa) CDR3s in the form of variability profiles (frequency distributions) and matrix showing residue positions and potential diversity. Vkappa CDR3 length sizes 8-9 according to Contact CDR definition.

Figure 22 shows the generated sequence diversity of exemplary synthetic light chain (Vlambda) CDR1s in the form of variability profiles (frequency distributions) and matrix showing residue positions and potential diversity. V_l CDR1 length sizes 9, 10 and 7 according to Contact CDR definition.

Figure 23 shows the generated sequence diversity of exemplary synthetic Vlambda 1, Vlambda 2 and Vlambda 3 light chain CDR2s in the form of variability profiles (frequency distributions) and matrix showing residue positions and potential diversity. Vlambda 1, Vlambda 2 and Vlambda 3 CDR2 length size 10 according to Contact CDR definition.

Figure 24 shows Vlambda CDR3 length distribution of sizes 8 to 11 amino acids which cover about 90% of the available CDR space. A separated analysis was performed for each length (see Fig. 25). VH CDR3 sizes are according to Contact CDR definition

Figure 25 shows the generated sequence diversity of exemplary synthetic light chain Vlambda CDR3s in the form of variability profiles (frequency distributions) and matrix showing residue positions and potential diversity. Vlambda CDR3 length sizes 8, 9, 10 and 11 according to Contact CDR definition.

Figure 26 shows the incorporation of CDR diversity into an expression template using Kunkel mutagenesis.

Figure 27 shows the sequence diversity of each light chain CDR as well as and the combined light chain library diversity. The number of variable positions and CDR sizes are according to Kabat definition.

Figure 28 shows the construction of the light chain library using a combination of overlapping nondegenerate and degenerate oligonucleotides which can be converted to double-stranded nucleic acids using single overlap extension polymerase chain reaction (SOE-PCR).

Figure 29 shows the affinity maturation of a test antibody (left panel) and resultant sequence diversity obtained (bottom panel) and improved binding of several representative clones (right panel).

Figure 30 shows a comparison of three CDR definitions: Kabat (Kabat *et al.*), Chothia (Chothia *et al.*), Contact considerations (MacCallum *et al.*) for VH (A) and VL (B) chains. The small triangles on the CDR segments point to locations where the insertions occur. Below the two graphs the number of amino acids of each CDR is displayed (*i.e.*, CDR lengths).

Figure 31 shows a sample of VBASE segments parsed and stored in FR1-CDR1-FR2-CDR2-FR3 format. CDR locations are identified according to Contact definition (MacCallum *et al.*) and the numbering scheme is according to Chothia (Chothia *et al.*). From these datasets individual data for each FR or CDR can be extracted.

Figure 32 shows the VBASE VH germline sequences stored as FR123 in FASTA format. The length of frameworks refers to the Contact CDR definition.

Figure 33 shows a hierarchical tree obtained from VBASE V_H segments in a FR123 format. UPGMA clustering algorithm has been used with the distance matrix computed using the p-distance (fraction of mismatches).

Figure 34 shows a hierarchical tree obtained from VBASE Vkappa segments in a FR123 format. UPGMA clustering algorithm has been used with the distance matrix computed using the p-distance (fraction of mismatches).

Figure 35 shows a hierarchical tree obtained from VBASE Vlambda segments in a FR123 format. UPGMA clustering algorithm has been used with the distance matrix computed using the p-distance (fraction of mismatches).

Figure 36 shows a Kabat VH input dataset (FR123 format) of known anti-protein antibodies visualized as a hierarchical tree (UPGMA).

Figure 37 shows a Kabat Vkappa input dataset (FR123 format) of known anti-protein antibodies visualized as a hierarchical tree (UPGMA).

Figure 38 shows a Kabat Vlambda input dataset (FR123 format) of known anti-protein antibodies visualized as a hierarchical tree (UPGMA).

Figure 39 shows the amino acid sequences of the 12 selected germline segments. In this figure germline CDR sequences are also included. These are replaced by mutagenized sequences as described in the following section.

Figure 40 shows a tree of frameworks appropriate for the antigen class of polysaccharides.

Figure 41 shows a flow chart depicting CDR1 and CDR2 variability profile selection. [See Example 3]

Figure 42 shows a flow chart depicting CDR3 variability profile selection. [See Example 3].

Figure 43 shows the selection process for subgroup pools of heavy chain sequences, subclass partitioning of subgroup sequences, and further partitioning of subclass populations on the basis of canonical structure.

Figure 44 shows a histogram of amino acid residue prevalence at each position within a VH-1_CDR1_6 population.

Figure 45 shows a histogram of amino acid residue prevalence at each position within a VH-1_CDR1_6_CS1 (canonical structure 2) population.

Figure 46 shows a histogram of amino acid residue prevalence at each position within a VH-1_CDR1_6_CS1-2 population.

Figure 47 shows a histogram of amino acid residue prevalence at each position within a VH-1_CDR1_6_CS1-3 population.

Figure 48 shows a histogram of amino acid residue prevalence at each position within a VH-1_CDR2_13 population.

Figures 49A and 49B show histograms of amino acid residue prevalence at each position within a VH-1_CDR2_13_CS2 population and a VH-1_CDR2_13_CS3 population, respectively.

Figures 50A and 50B show histograms of amino acid residue prevalence at each position within a VH-1_CDR2_13_CS2-1 population and a VH-1_CDR2_13_CS3-1 population, respectively.

Figures 51A, 51B and 51C show histograms of amino acid residue prevalence at each position within a VH_CDR3-9 population, VH_CDR3-15 population, and a VH_CDR3-18 population, respectively.

Detailed Description of the Invention

In order to provide a clear understanding of the specification and claims, the following definitions are provided below.

Definitions

As used herein the term "antibody binding regions" refers to one or more portions of an immunoglobulin or antibody variable region capable of binding an antigen(s). Typically, the antibody binding region is, for example, an antibody light chain (VL) (or variable region thereof), an antibody heavy chain (VH) (or variable region thereof), a heavy chain Fd region, a combined antibody light and heavy chain (or variable region thereof) such as a Fab, $F(ab')_2$, single domain, or single chain antibody (scFv), or a full length antibody, for example, an IgG (*e.g.*, an IgG1, IgG2, IgG3, or IgG4 subtype), IgA1, IgA2, IgD, IgE, or IgM antibody.

The term "framework region" refers to the art recognized portions of an antibody variable region that exist between the more divergent CDR regions. Such framework regions are typically referred to as frameworks 1 through 4 (FR1, FR2, FR3, and FR4) and provide a scaffold for holding, in three-dimensional space, the three CDRs found in a heavy or light chain antibody variable region, such that the CDRs can form an antigen-binding surface.

The term "threshold frequency of occurrence" refers to a criterion of the invention which requires that a selected sequence for use in the universal antibody library be derived from a sequence which has been determined to be a sequence favored to be expressed by immune cells when, for example, responding to a particular class of antigens. Typically, such expressed (rearranged) sequences determined to meet the threshold frequency of occurrence are sequences which are expressed at a percent occurrence of about 10% or more.

The term "threshold frequency of germline origin" refers to a criterion of the invention which requires that a selected sequence (*i.e.*, expressed or rearranged sequence) for use in the universal antibody library be derived from a sequence which has been determined to be a germline sequence favored to be expressed by immune cells when, for example, responding to a particular class of antigens. Typically, sequences determined to meet the threshold frequency of germline origin are sequences which are derived or originate from a germline sequence at a percent occurrence of about 10% or more.

The term "predetermined antigen class", or "class of antigens" or "antigen class" refers to antigens which are structurally / chemically similar in terms of their basic composition. Typical antigen classes are proteins (polypeptides), peptides, polysaccharides, polynucleotides, and small molecules.

The term "canonical structure" includes considerations as to the linear sequence of the antibody, for example, as catalogued in the Kabat database. The Kabat numbering scheme is a widely adopted standard for numbering the amino acid residues of an antibody variable domain in a consistent manner. Additional structural considerations, for example, those differences not fully reflected by Kabat numbering, for example, as described by Chothia *et al.* and revealed by, for example, crystallography and three-dimensional modeling, can also be used to determine the canonical structure of an antibody. Accordingly, a given antibody sequence may be placed into a canonical class which allows for, among other things, identifying appropriate acceptor sequences. Kabat numbering of antibody amino acid sequence and structural considerations, for example, as described by Chothia *et al.*, and its implication for construing canonical aspects of a given antibody, are described in the literature (see also, *e.g.*, Materials and Methods, below). The term "canonical structure" also refers to the main chain

conformation that is adopted by one of the antigen binding loops. From comparative structural comparisons, it has been found that five of the six antigen binding loops only have a limited repertoire of available conformations. Each canonical structure can be characterized by the polypeptide backbone torsion angles. Correspondent loops between antibodies may therefore have very similar three dimensional structures despite high amino acid sequence variability in most parts of the loops (Chothia and Lesk, 1987 *J. Mol. Biol.* 196, 901-917 Chothia *et al.*, 1989 *Nature* 342, 877-883 Martin and Thornton, 1996 *J. Mol. Biol.* 263, 800-815). Furthermore, there is a relationship between the adopted loop structure and amino acid sequences surrounding it. The conformation of a particular canonical class is determined by the length of the loop and amino acid residues at key positions, interacting within the loop and outside in the conserved framework. These key amino acids often interact through hydrogen bonding. Assignment to a particular canonical class can therefore be made based on the presence of these key amino acid residues.

The term "defined CDR region" refers to a complementarity determining region (CDR) of which three make up the binding character of a light chain variable region and/or heavy chain variable region of a binding molecule. There are three CDRs in each of the variable heavy and variable light sequences designated CDR1, CDR2 and CDR3, for each of the variable regions. Defined CDR regions contribute to the functional activity of an antibody molecule and may be separated by amino acid sequences that are merely scaffolding or framework regions. The exact definitional CDR boundaries and lengths are subject to different classification systems. CDRs may therefore be referred to by Kabat, Chothia, contact or any other boundary definitions. Despite differing boundaries, they all have some overlapping residues in what constitute the so called "hypervariable regions" within the variable sequences. These CDR definitions will therefore differ in length and boundary areas with respect to the adjacent framework region. See for example Kabat , Chothia, and/or MacCallum *et al.*, (see, *e.g.,* Kabat *et al.*, *In* "Sequences of Proteins of Immunological Interest," U.S. Department of Health and Human Services, 1983; Chothia *et al.*, *J. Mol. Biol.* 196:901-917, 1987; and MacCallum *et al.*, *J. Mol. Biol.* 262:732-745 (1996); the contents of which are incorporated herein in their entirety).

The term "conserved amino acid residue" refers to an amino acid residue determined to occur with a frequency between germ line sequence and CDR sequence or between CDRs of a given canonical class and/or length, that is high, typically at least 50% or more (e.g., at about 60%, 70%, 80%, 90%, 95%, or higher), for a given residue position. When a given residue is determined to occur at such a high frequency, it is determined to be conserved and thus represented in the libraries of the invention as a "fixed" or "constant" residue, at least for that amino acid residue position in the CDR

region being analyzed. Typically, no nucleic acid mutagenesis/variability is introduced for a conserved amino acid (codon) position, but rather, the residue is fixed and predetermined.

The term "semi-conserved amino acid residue" refers to amino acid residues determined to occur with a frequency between germ line sequence and CDR sequence or between CDRs of a given canonical class and/or length that is high, for 2 to 3 residues for a given residue position. When 2-3 residues, preferably 2 residues, that together, are represented at a frequency of about 40% of the time or higher (e.g., 50%, 60%, 70%, 80%, 90% or higher), the residues are determined to be semi-conserved and thus represented in the libraries of the invention as a "semi-fixed" at least for that amino acid residue position in the CDR region being analyzed. Typically, an appropriate level of nucleic acid mutagenesis/variability is introduced for a semi-conserved amino acid (codon) position such that the 2 to 3 residues are properly represented. Thus, each of the 2 to 3 residues can be said to be "semi-fixed" for this position.

The term "variable amino acid residue" refers to amino acid residues determined to occur with a frequency between germ line sequence and CDR sequence or between CDRs of a given canonical class and/or length that is variable for a given residue position. When many residues appear at a given position, the residue position is determined to be variable and thus represented in the libraries of the invention as variable at least for that amino acid residue position in the CDR region being analyzed. Typically, an appropriate level of nucleic acid mutagenesis/variability is introduced for a variable amino acid (codon) position such that an accurate spectrum of residues are properly represented. Of course, it is understood that, if desired, the consequences or variability of any amino acid residue position, i.e., conserved, semi-conserved, or variable, can be represented, explored or altered using, as appropriate, any of the mutagenesis methods disclosed herein, e.g., LTM, WTM, WTM with doping, and/or extended WTM..

The term "variability profile" refers to the cataloguing of amino acids and their respective frequency rates of occurrence present at a particular CDR position. The CDR positions are derived from an aligned CDR dataset grouped according to desired characteristics. At each CDR position, ranked amino acid frequencies are added to that position's variability profile until the amino acids' combined frequencies reach a predetermined "high" threshold value.

The term "amino acid" or "amino acid residue" typically refers to an amino acid having its art recognized definition such as an amino acid selected from the group consisting of: alanine (Ala); arginine (Arg); asparagine (Asn); aspartic acid (Asp); cysteine (Cys); glutamine (Gln); glutamic acid (Glu); glycine (Gly); histidine (His); isoleucine (Ile): leucine (Leu); lysine (Lys); methionine (Met); phenylalanine (Phe);

proline (Pro); serine (Ser); threonine (Thr); tryptophan (Trp); tyrosine (Tyr); and valine (Val) although modified, synthetic, or rare amino acids may be used as desired.. Generally, amino acids can be grouped as having a nonpolar sidechain (*e.g.*, Ala, Cys, Ile, Leu, Met, Phe, Pro, Val); a negatively charged side chain (*e.g.*, Asp, Glu); a positively charged sidechain (*e.g.*, Arg, His, Lys); or an uncharged polar sidechain (*e.g.*, Asn, Cys, Gln, Gly, His, Met, Phe, Ser, Thr, Trp, and Tyr).

The term "library" refers to two or more antibody molecules (or fragments thereof) having a diversity as described herein mutagenized according to the method of the invention. The antibodies of the library can be in the form of polynucleotides, polypeptides, polynucleotides and polypeptides, polynucleotides and polypeptides in a cell free extract, or as polynucleotides and/or polypeptides in the context of a phage, prokaryotic cells, or in eukaryotic cells.

The term "polynucleotide(s)" refers to nucleic acids such as DNA molecules and RNA molecules and analogs thereof (*e.g.*, DNA or RNA generated using nucleotide analogs or using nucleic acid chemistry). As desired, the polynucleotides may be made synthetically, *e.g.*, using art-recognized nucleic acid chemistry or enzymatically using, *e.g.*, a polymerase, and, if desired, be modified. Typical modifications include methylation, biotinylation, and other art-known modifications. In addition, the nucleic acid molecule can be single-stranded or double-stranded and, where desired, linked to a detectable moiety.

The term "mutagenesis" refers to, unless otherwise specified, any art recognized technique for altering a polynucleotide or polypeptide sequence. Preferred types of mutagenesis include walk-through mutagenesis (WTM), beneficial walk-through mutagenesis, look-through mutagenesis (LTM), improved look-through mutagenesis (LTM2), WTM using doped nucleotides for achieving codon bias, extended WTM for holding short regions of sequence as constant or fixed within a region of greater diversity, or combinations thereof.

The term "combinatorial beneficial mutagenesis" refers to a combination library of coding sequences that encode degenerate mixtures of V_L and/or V_H CDR amino-acid sequence variations initially identified from the predetermined LTM amino acid mutagenesis screen as having an alteration on a measurable property. In the combinatorial beneficial mutation approach, oligonucleotide coding sequences are generated which represent combinations of these beneficial mutations identified by LTM. These combinations may be combinations of different beneficial mutations within a single CDR, mutations within two or more CDRs within a single antibody chain, or mutations within the CDRs of different antibody chains.

Detailed Description

Overview

Antibodies are powerful diagnostic and therapeutic tools. Antibody libraries comprising candidate binding molecules that can be readily screened against targets are desirable. The full promise of a comprehensive universal antibody library has remained elusive. Synthetic libraries suffer from noise and too much diversity that is not naturally occurring. Entirely human libraries are biased against certain antigen classes and only as diverse as capture techniques allow for. The present invention provides a universal antibody library that is comprehensive and can be readily screened using, for example, high throughput methods to obtain new therapeutics.

In particular, the universal antibody library (UAL) has the potential to recognize any antigen. Other significant advantages of the library include greater diversity, for example, to self antigens that are usually lost in a expressed human library because self reactive antibodies are removed by the donor's immune system by negative selection. Another feature is that screening the universal antibody library (UAL) using positive clone selection by FACS (florescence activated cell sorter) bypasses the standard and tedious methodology of generating a hybridoma library and supernatant screening. Still further, the UAL library can be re-screened to discover additional antibodies against other desired targets.

1.1 Identifying and Selecting Universal Antibody Components Using Bioinformatics

The first step in building a universal antibody library (UAL) of the invention is selecting sequences that meet certain predetermined criteria. For example, the Kabat database, a electronic database containing non-redundant rearranged antibody sequences can be queried for those sequences that are most frequently represented, in particular, against a particular antigen class. The antigen class can include, for example, protein and peptide antigens but also small molecules, polysaccharides, and polynucleotides. A clustering analysis of the framework sequences of these antibodies is performed followed by a comparison (using the BLAST search algorithm) with germline sequences (V BASE database) to determine the most frequently used germline families that subsequently rearrange to generate functional antibodies that recognize a given antigen class, for example, proteinaceous antigens or targets.

The candidate framework sequences that represent the largest and most structurally diverse groups of functional antibodies are then chosen, and the canonical structures of CDR1 and CDR2 are then determined, to determine the length of the CDRs and thus, the diversity that can be accommodated within the frameworks. For CDR3, a size distribution of lengths is performed to identify a frequency analysis of rearranged antibody sequences.

The method for deriving amino acid sequences of the CDRs includes a frequency analysis and the generation of the corresponding variability profiles (VP) of existing rearranged antibody sequences. Invariant positions are fixed while the highest frequency amino acids are chosen as wildtype at other positions. These wildtype amino acids are then systematically altered using, mutagenesis, *e.g.* walk-through mutagenesis (WTM), to generate the universal antibody library.

The universal library construction strategy involves selection of framework sequences followed by design of the hypervariable CDR loops. For framework sequence selection, a subset of all available framework scaffolds determined to have been expressed in response to a particular antigen are arrayed. By determining the frameworks that are most frequently expressed in nature in response to a given antigen class an appropriate framework acceptor is selected. For example, to determine the preferred acceptor frameworks expressed in response to protein-based antigens, the Kabat database (accessible at http://www.kabatdatabase.com) is searched for "protein-directed" frameworks. If preferred acceptor sequences are needed for presenting CDRs against a different antigen class, and/or, acceptor sequences of a particular species, the Kabat protein sequence filter is set accordingly. For example, to determine sequences for use as human therapeutics against protein-based targets, the filter is set to focus only on human antibody sequences (not mouse, rat, or chicken sequences, *etc.*) that recognize protein/peptide antigens. This greatly reduces redundancy in the dataset and sequence information that would bias results.

The above step minimizes the need to generate numerous different synthetic framework scaffolds and typically results in a data set of potential acceptors of about 600 sequences or less. Accordingly, the resultant number of sequences is easily manageable for further analysis to determine the germline precursor sequences that give rise to the rearranged gene sequences that are selected by antigen class. This second determination of germline origin refines the selection of the antibody sequences that have been selected by an antigen class because it identifies if there are optimal (or high frequency) germline framework sequences that are overrepresented. Indeed, it has been observed that in some polyclonal responses against certain antigens, where a large number of rearranged antibody sequence are produced, that only a few acceptor framework sequences are used. In such a case, the antibody sequence and binding diversity for the antigens is chiefly localized to the CDRs not the frameworks. The above bioinformatic analysis focuses on V_H genes for descriptive purposes, but it will be understood that genes for both Vλ and Vκ are similarly evaluated.

1.2 Design Strategies for Maximizing CDR Diversity

The choice of candidate frameworks based on the criteria of the invention dictates both the CDR sizes to be introduced and the initial amino acid sequence diversity. When the antibody sequences are identified for 1) frequency of occurrence against an antigen class and 2) germline frequency, the sequences can then be arrayed according to their canonical class. The canonical class is determined using the conventions as described by Chothia (see Materials and Methods, below). Of a given set of antibody sequences, the majority of the antibody sequences identified may fall with in a certain canonical class. The canonical class then dictates the number of amino acid residues that can be accommodated in the CDRs. For example, if the canonical class is 1-3, then CDR1 would have a 6 amino acid loop and CDR2 would have a 13 amino acid loop. For the heavy chain variable sequence the J segment sequence contribution is relatively well conserved such that typically, only the best fit sequence from a subset of only six sequences need be considered. A CDR amino acid frequency analysis of the Kabat and V BASE databases allows identification of CDR amino acid residue positions that fall within two categories, e.*g.*, 1) positions that should be conserved, and 2) positions that are suitable for diversity generation.

In designing VH-CDR2, diversity analysis in the V BASE and Kabat databases is approached in an similar manner as was performed for VH-CDR1 above.

In designing V_H CDR3 diversity, CDR3 sequences of antibodies from the Kabat database are aligned according to their size and antigen class. Lengths of CDR3s of antibodies recognizing non-protein and protein/peptide antigens are compared and a frequency analysis is performed and a threshold frequency of 10% is used to identify the most favorable sequences to be used in designating the CDR3 diversity. Because CDR3 size and amino acid residue frequency analysis is performed using, *e.g.*, the immunoglobulin (D) and J gene rearranged sequences, there are no "CDR3" germline equivalents for direct filtered Kabat and V BASE comparisons. However, a filtered Kabat frequency analysis or variability profile (VP) can be generated (Figures 11 and 12) for each rearranged CDR3 size can be performed which reveals, for each size classification, the most frequent amino acid throughout the CDR3 positions and results in a consensus "wild type" sequence. Surprisingly, this "consensus" or "frequency" approach identifies those particular amino acids under high selective pressure. Accordingly, these residue position are typically fixed with diversity being introduced into remaining amino acid positions (taking into account the identified preference for certain amino acids to be present at these positions).

When designing the diversity for any of the above-mentioned CDRs, modified amino acid residues, for example, residues outside the traditional 20 amino acids used in most polypeptides, *e.g.*, homocysteine, can be incorporated into the CDRs as desired.

This is carried out using art recognized techniques which typically introduce stop codons into the polynucleotide where the modified amino acid residue is desired. The technique then provides a modified tRNA linked to the modified amino acid to be incorporated (a so-called suppressor tRNA of, *e.g.*, the stop codon amber, opal, or ochre) into the polypeptide (see, *e.g.*, Köhrer *et al.*, Import of amber and ochre suppressors tRNAs into mammalian cells: A general approach to site-specific insertion of amino acid analogues into proteins, *PNAS*, 98, 14310-14315 (2001)).

2. Computer-Assisted Universal Antibody Library (UAL) Construction

The universal antibody libraries of the invention and their construction is conducted with the benefit of sequence and structural information concerning the antibody diversity to be generated, such that the potential for generating improved antibodies is increased. Modeling information can also be used to guide the selection of amino acid diversity to be introduced into the defined regions, *e.g.*, CDRs. Still further, actual results obtained with the antibodies of the invention can guide the selection (or exclusion), *e.g.*, affinity maturation, of subsequent antibodies to be made and screened in an iterative manner.

In a particular embodiment, *in silico* modeling is used to eliminate the production of any antibodies predicted to have poor or undesired structure and/or function. In this way, the number of antibodies to be produced can be sharply reduced thereby increasing signal-to-noise in subsequent screening assays. In another particular embodiment, the *in silico* modeling is continually updated with additional modeling information, from any relevant source, *e.g.*, from gene and protein sequence and three-dimensional databases and/or results from previously tested antibodies, so that the *in silico* database becomes more precise in its predictive ability (Fig. 1).

In yet another embodiment, the *in silico* database is provided with the assay results, *e.g.*, binding affinity / avidity of previously tested antibodies and categorizes the antibodies, based on the assay criterion or criteria, as responders or nonresponders, *e.g.*, as antibodies that bind well or not so well. In this way, the affinity maturation of the invention can equate a range of functional responses with particular sequence and structural information and use such information to guide the production of future antibodies to be tested. The method is especially suitable for screening antibody or antibody fragments for a particular binding affinity to a target antigen using, *e.g.*, a Biacore assay.

Accordingly, mutagenesis of noncontiguous residues within a region can be desirable if it is known, *e.g.*, through in silico modeling, that certain residues in the region will not participate in the desired function. The coordinate structure and spatial interrelationship between the defined regions, *e.g.*, the functional amino acid residues in

the defined regions of the antibody, *e.g.*, the diversity that has been introduced, can be considered and modeled. Such modeling criteria include, *e.g.*, amino acid residue side group chemistry, atom distances, crystallography data, etc. Accordingly, the number antibodies to be produced can be intelligently minimized.

In a preferred embodiment, one or more of the above steps are computer-assisted. In a particular embodiment, the computer assisted step comprises, *e.g.*, mining the Kabat database and, optionally, cross-referencing the results against Vbase, whereby certain criteria of the invention are determined and used to design the desired CDR diversity (Figs. 1-2). The method is also amenable to being carried out, in part or in whole, by a device, *e.g.*, a computer driven device. For example, database mining antibody sequence selection, diversity design, oligonucleotide synthesis, PCR-mediated assembly of the foregoing, and expression and selection of candidate antibodies that bind a given target, can be carried out in part or entirely, by interlaced devices. In addition, instructions for carrying out the method, in part or in whole, can be conferred to a medium suitable for use in an electronic device for carrying out the instructions. In sum, the methods of the invention are amendable to a high throughput approach comprising software (*e.g.*, computer-readable instructions) and hardware (*e.g.*, computers, robotics, and chips).

3. Synthesizing Universal Antibody Libraries

In one embodiment, the universal antibody libraries (UAL) of the invention are generated for screening by synthesizing individual oligonucleotides that encode the defined region of the polypeptide and have no more than one codon for the predetermined amino acid. This is accomplished by incorporating, at each codon position within the oligonucleotide either the codon required for synthesis of the wild-type polypeptide or a codon for the predetermined amino acid and is referred to as look-through mutagenesis (LTM) (see, *e.g.*, U.S.S.N. 60/483282).

In another embodiment, when diversity at multiple amino acid positions is required, walk-through mutagenesis (WTM) can be used (see *e.g.*, U.S. Pat Nos. 6,649,340; 5,830,650; and 5,798,208; and U.S.S.N. 60/483,282. WTM allows for multiple mutations to be made with a minimum number of oligonucleotides. The oligonucleotides can be produced individually, in batches, using, *e.g.*, doping techniques, and then mixed or pooled as desired.

The mixture of oligonucleotides for generation of the library can be synthesized readily by known methods for DNA synthesis. The preferred method involves use of solid phase beta-cyanoethyl phosphoramidite chemistry (*e.g.*, see U.S. Pat. No. 4,725,677). For convenience, an instrument for automated DNA synthesis can be used containing specified reagent vessels of nucleotides. The polynucleotides may also be

synthesized to contain restriction sites or primer hybridization sites to facilitate the introduction or assembly of the polynucleotides representing, *e.g.*, a defined region, into a larger gene context.

The synthesized polynucleotides can be inserted into a larger gene context, *e.g.*, a single chain antibody (scFv) using standard genetic engineering techniques. For example, the polynucleotides can be made to contain flanking recognition sites for restriction enzymes (*e.g.*, see U.S. Pat. No. 4,888,286). The recognition sites can be designed to correspond to recognition sites that either exist naturally or are introduced in the gene proximate to the DNA encoding the region. After conversion into double stranded form, the polynucleotides are ligated into the gene or gene vector by standard techniques. By means of an appropriate vector (including, *e.g.*, phage vectors, plasmids) the genes can be introduced into a cell-free extract, phage, prokaryotic cell, or eukaryotic cell suitable for expression of the antibodies.

Alternatively, partially overlapping polynucleotides, typically about 20-60 nucleotides in length, are designed. The internal polynucleotides are then annealed to their complementary partner to give a double-stranded DNA molecule with single-stranded extensions useful for further annealing. The annealed pairs can then be mixed together, extended, and ligated to form full-length double-stranded molecules using PCR (see, *e.g.*, Example 3). Convenient restriction sites can be designed near the ends of the synthetic gene for cloning into a suitable vector. The full-length molecules can then be ligated into a suitable vector.

When partially overlapping polynucleotides are used in the gene assembly, a set of degenerate nucleotides can also be directly incorporated in place of one of the polynucleotides. The appropriate complementary strand is synthesized during the extension reaction from a partially complementary polynucleotide from the other strand by enzymatic extension with a polymerase. Incorporation of the degenerate polynucleotides at the stage of synthesis also simplifies cloning where more than one domain or defined region of a gene is mutagenized or engineered to have diversity.

In another approach, the antibody is present on a single stranded plasmid. For example, the gene can be cloned into a phage vector or a vector with a filamentous phage origin of replication that allows propagation of single-stranded molecules with the use of a helper phage. The single-stranded template can be annealed with a set of degenerate polynucleotides representing the desired mutations and elongated and ligated, thus incorporating each analog strand into a population of molecules that can be introduced into an appropriate host (see, *e.g.*, Sayers, J. R. *et al.*, Nucleic Acids Res. 16: 791-802 (1988)). This approach can circumvent multiple cloning steps where multiple domains are selected for mutagenesis.

Polymerase chain reaction (PCR) methodology can also be used to incorporate polynucleotides into a gene, for example, CDR diversity into framework regions. For example, the polynucleotides themselves can be used as primers for extension. In this approach, polynucleotides encoding the mutagenic cassettes corresponding to the defined region (or portion thereof) are complementary to each other, at least in part, and can be extended to form a large gene cassette (*e.g.*, a scFv) using a polymerase, *e.g.*, using PCR amplification.

The size of the library will vary depending upon the CDR length and the amount of CDR diversity which needs to be represented using, *e.g.*, WTM or LTM. Preferably, the library will be designed to contain less than 10^{15}, 10^{14}, 10^{13}, 10^{12}, 10^{11}, 10^{10}, 10^{9}, 10^{8}, 10^{7}, and more preferably, 10^{6} antibodies or less.

The description above has centered on representing antibody diversity by altering the polynucleotide that encodes the corresponding polypeptide. It is understood, however, that the scope of the invention also encompasses methods of representing the antibody diversity disclosed herein by direct synthesis of the desired polypeptide regions using protein chemistry. In carrying out this approach, the resultant polypeptides still incorporate the features of the invention except that the use of a polynucleotide intermediate can be eliminated.

For the libraries described above, whether in the form of polynucleotides and/or corresponding polypeptides, it is understood that the libraries may be also attached to a solid support, such as a microchip, and preferably arrayed, using art recognized techniques.

The method of this invention is especially useful for modifying candidate antibody molecules by way of affinity maturation. Alterations can be introduced into the variable region and/or into the framework (constant) region of an antibody. Modification of the variable region can produce antibodies with better antigen binding properties, and, if desired, catalytic properties. Modification of the framework region can also lead to the improvement of chemo-physical properties, such as solubility or stability, which are especially useful, for example, in commercial production, bioavailabilty, and affinity for the antigen. Typically, the mutagenesis will target the Fv region of the antibody molecule, *i.e.*, the structure responsible for antigen-binding activity which is made up of variable regions of two chains, one from the heavy chain (VH) and one from the light chain (VL). Once the desired antigen-binding characteristics are identified, the variable region(s) can be engineered into an appropriate antibody class such as IgG, IgM, IgA, IgD, or IgE. In a preferred embodiment, an identified candidate binding molecule is subjected to affinity maturation to increase the affinity / avidity of the binding molecule to a target/antigen.

4. Expression and Screening Systems

Libraries of polynucleotides generated by any of the above techniques or other suitable techniques can be expressed and screened to identify antibodies having desired structure and/or activity. Expression of the antibodies can be carried out using cell-free extracts (and *e.g.*, ribosome display), phage display, prokaryotic cells, or eukaryotic cells (*e.g.*, yeast display).

In one embodiment, the polynucleotides are engineered to serve as templates that can be expressed in a cell free extract. Vectors and extracts as described, for example in U.S. Patent Nos. 5,324,637; 5,492,817; 5,665,563, can be used and many are commercially available. Ribosome display and other cell-free techniques for linking a polynucleotide (*i.e.*, a genotype) to a polypeptide (*i.e.*, a phenotype) can be used, *e.g.*, Profusion™ (see, *e.g.*, U.S. Patent Nos. 6,348,315; 6,261,804; 6,258,558; and 6,214,553).

Alternatively, the polynucleotides of the invention can be expressed in a convenient *E. coli* expression system, such as that described by Pluckthun and Skerra. (Pluckthun, A. and Skerra, A., Meth. Enzymol. 178: 476-515 (1989); Skerra, A. *et al.*, Biotechnology 9: 273-278 (1991)). The mutant proteins can be expressed for secretion in the medium and/or in the cytoplasm of the bacteria, as described by M. Better and A. Horwitz, Meth. Enzymol. 178: 476 (1989). In one embodiment, the single domains encoding VH and VL are each attached to the 3' end of a sequence encoding a signal sequence, such as the ompA, phoA or pelB signal sequence (Lei, S. P. *et al.*, J. Bacteriol. 169: 4379 (1987)). These gene fusions are assembled in a dicistronic construct, so that they can be expressed from a single vector, and secreted into the periplasmic space of *E. coli* where they will refold and can be recovered in active form. (Skerra, A. *et al.*, Biotechnology 9: 273-278 (1991)). For example, antibody heavy chain genes can be concurrently expressed with antibody light chain genes to produce antibody or antibody fragments.

In another embodiment, the antibody sequences are expressed on the membrane surface of a prokaryote, *e.g.*, *E. coli*, using a secretion signal and lipidation moiety as described, *e.g.*, in US20040072740A1; US20030100023A1; and US20030036092A1.

In still another embodiment, the polynucleotides can be expressed in eukaryotic cells such as yeast using, for example, yeast display as described, *e.g.*, in U.S. Patent Nos. 6,423,538; 6,331,391; and 6,300,065. In this approach, the antibodies of the library (*e.g.*, scFvs) are fused to a polypeptide that is expressed and displayed on the surface of the yeast.

Higher eukaryotic cells for expression of the antibodies of the invention can also be used, such as mammalian cells, for example myeloma cells (*e.g.*, NS/0 cells), hybridoma cells, or Chinese hamster ovary (CHO) cells. Typically, the antibodies when

expressed in mammalian cells are designed to be expressed into the culture medium, or expressed on the surface of such a cell. The antibody or antibody fragments can be produced, for example, as entire antibody molecules or as individual VH and VL fragments, Fab fragments, single domains, or as single chains (sFv) (see *e.g.*, Huston, J. S. *et al.*, Proc. Natl. Acad. Sci. USA 85: 5879-5883 (1988)).

The screening of the expressed antibodies (or antibodies produced by direct synthesis) can be done by any appropriate means. For example, binding activity can be evaluated by standard immunoassay and/or affinity chromatography. Screening of the antibodies of the invention for catalytic function, *e.g.*, proteolytic function can be accomplished using a standard hemoglobin plaque assay as described, for example, in U.S. Patent No. 5,798,208. Determining the ability of candidate antibodies to bind therapeutic targets can be assayed *in vitro* using, *e.g.*, a Biacore instrument, which measures binding rates of an antibody to a given target or antigen. *In vivo* assays can be conducted using any of a number of animal models and then subsequently tested, as appropriate, in humans.

Exemplification

Throughout the examples, the following materials and methods were used unless otherwise stated.

Materials and Methods

In general, the practice of the present invention employs, unless otherwise indicated, conventional techniques of chemistry, molecular biology, recombinant DNA technology, PCR technology, immunology (especially, *e.g.*, antibody technology), expression systems (*e.g.*, cell-free expression, phage display, ribosome display, and Profusion™), and any necessary cell culture that are within the skill of the art and are explained in the literature. See, *e.g.*, Sambrook, Fritsch and Maniatis, *Molecular Cloning: Cold Spring Harbor Laboratory Press* (1989); *DNA Cloning*, Vols. 1 and 2, (D.N. Glover, Ed. 1985); *Oligonucleotide Synthesis* (M.J. Gait, Ed. 1984); *PCR Handbook Current Protocols in Nucleic Acid Chemistry*, Beaucage, Ed. John Wiley & Sons (1999) (Editor); *Oxford Handbook of Nucleic Acid Structure*, Neidle, Ed., Oxford Univ Press (1999); *PCR Protocols: A Guide to Methods and Applications*, Innis *et al.*, Academic Press (1990); *PCR Essential Techniques: Essential Techniques*, Burke, Ed., John Wiley & Son Ltd (1996); *The PCR Technique: RT-PCR,* Siebert, Ed., Eaton Pub. Co. (1998); *Antibody Engineering Protocols (Methods in Molecular Biology),* 510, Paul, S., Humana Pr (1996); *Antibody Engineering: A Practical Approach (Practical Approach Series, 169)*, McCafferty, Ed., Irl Pr (1996); *Antibodies: A Laboratory Manual,* Harlow *et al.*, C.S.H.L. Press, Pub. (1999); *Current Protocols in Molecular Biology*, eds. Ausubel *et al.*, John Wiley & Sons (1992); *Large-Scale Mammalian Cell Culture Technology,* Lubiniecki, A., Ed., Marcel Dekker, Pub., (1990). *Phage Display : A Laboratory Manual*, C. Barbas (Ed.), CSHL Press, (2001); *Antibody Phage Display*, P O'Brien (Ed.), Humana Press (2001); Border *et al.*, Yeast surface display for screening combinatorial polypeptide libraries, *Nature Biotechnology*,15(6):553-7 (1997); Border *et al.*, Yeast surface display for directed evolution of protein expression, affinity, and stability, *Methods Enzymol.*, 328:430-44 (2000); ribosome display as described by Pluckthun *et al.* in U.S. Patent No. 6,348,315, and Profusion™ as described by Szostak *et al.* in U.S. Patent Nos. 6,258,558; 6,261,804; and 6,214,553;, and bacterial periplasmic expression as described in US20040058403A1.

Further details regarding antibody sequence analysis using Kabat conventions may be found, *e.g.*, in Johnson *et al.*, The Kabat database and a bioinformatics example, *Methods Mol Biol.* 2004;248:11-25; Johnson *et al.*, Preferred CDRH3 lengths for antibodies with defined specificities, *Int Immunol.* 1998, Dec;10(12):1801-5; Johnson *et al.*, SEQHUNT. A program to screen aligned nucleotide and amino acid sequences,

Methods Mol Biol. 1995;51:1-15. and Wu *et al.*, Length distribution of CDRH3 in antibodies; and Johnson *et al.*, *Proteins.* 1993 May;16(1):1-7. Review).

Further details regarding antibody sequence analysis using Chothia conventions may be found, *e.g.*, in Chothia *et al.*, Structural determinants in the sequences of immunoglobulin variable domain, *J Mol Biol.* 1998 May 1;278(2):457-79; Morea *et al.*, Antibody structure, prediction and redesign, *Biophys Chem.* 1997 Oct;68(1-3):9-16. ; Morea *et al.*, Conformations of the third hypervariable region in the VH domain of immunoglobulins; *J Mol Biol.* 1998 Jan 16;275(2):269-94; Al-Lazikani *et al.*, Standard conformations for the canonical structures of immunoglobulins, *J Mol Biol.* 1997 Nov 7;273(4):927-48. Barre *et al.*, Structural conservation of hypervariable regions in immunoglobulins evolution, Nat Struct Biol. 1994 Dec;1(12):915-20; Chothia *et al.*, Structural repertoire of the human VH segments, *J Mol Biol.* 1992 Oct 5;227(3):799-817 Conformations of immunoglobulin hypervariable regions, Nature. 1989 Dec 21-28;342(6252):877-83; and Chothia *et al.*, Review Canonical structures for the hypervariable regions of immunoglobulins, *J Mol Biol.* 1987 Aug 20;196(4):901-17).

Further details regarding Chothia analysis are described, for example, in Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol. 1998 Jan 16;275(2):269-94; Chothia C, Lesk AM, Gherardi E, Tomlinson IM, Walter G, Marks JD, Llewelyn MB, Winter G. Structural repertoire of the human VH segments. J Mol Biol. 1992 Oct 5;227(3):799-817; Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, *et al.* Conformations of immunoglobulin hypervariable regions. Nature. 1989 Dec 21-28;342(6252):877-83; Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol. 1987 Aug 20;196(4):901-17; and Chothia C, Lesk AM. The evolution of protein structures. Cold Spring Harb Symp Quant Biol. 1987;52:399-405.

Further details regarding CDR contact considerations are described, for example, in MacCallum RM, Martin AC, Thornton JM. Antibody-antigen interactions: contact analysis and binding site Topography. J Mol Biol. 1996 Oct 11;262(5):732-45.

Further details regarding the antibody sequences and databases referred to herein are found, *e.g.*, in Tomlinson IM, Walter G, Marks JD, Llewelyn MB, Winter G. The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops. J Mol Biol. 1992 Oct 5;227(3):776-98; Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 2001 Mar;17(3):282-3; [VBDB] www.mrc-cpe.cam.ac.uk/vbase-ok.php?menu=901; [KBTDB] www.kabatdatabase.com; [BLST] www.ncbi.nlm.nih.gov/BLAST/ [CDHIT] bioinformatics.ljcrf.edu/cd-hi/;

[EMBOSS] www.hgmp.mrc.ac.uk/Software/EMBOSS/; [PHYLIP] evolution.genetics.washington.edu/phylip.html; and [FASTA] fasta.bioch.virginia.edu.

Bacterial expression libraries are typically constructed as follows. The template sequence (Fig. 18) is cloned into an appropriate expression-display vector such as the APEx expression display system described in Harvey et al. PNAS 101 (25): 9193. (2004). The walkthrough and extended walkthrough libraries are prepared by Kunkel mutagenesis of the construct prepared with sequence incorporated into the APEx vector. A single-stranded template for Kunkel mutagenesis was prepared using standard protocols Sidhu, S. S. (2000) *Methods Enzymol.* 328:333-63. Kunkel mutagenesis of the template was carried out according to standard methods, as detailed, for example, in Kunkel, T. A. (1985) *Proc. Natl. Acad. Sci.* USA 82:488-92; Kunkel, T. A. et al. (1987) *Meth. Enzymol.* 154: 367-82; Zoller, M. J. and Smith, M. (1983) *Meth. Enzymol.* 100:468-500; Hanahan, D. (1983) *J. Mol. Biol.* 166:557-80; and Maniatis, T., Fritsch, E. F. and Sambrook, J. (1989) in Molecular Cloning, A Laboratory Manual.

Fig. 26 shows general steps in the Kunkel mutagenesis process for introducing a collection of CDR library oligonucleotides into a template antibody coding sequence. Initially, the single-stranded uracylated template is reacted with a collection of L3 oligonculeotides (green fragment) that carries the selected codon substitutions for CDRL3. For template utilizing SEQID-1, there are two sizes of CDRL3: size 8 and size 9. For each of these size, each of the walkthrough amino acids will generate a single oligonucleotide mixture. Nine walkthrough oligonucleotide mixtures are shown for size 8 (VKIII_3_8*), and nine walkthrough mixtures for size 9 (VKIII_3_9*). All 18 oligonucleotide mixtures are combined in an equimolar fashion and Kunkel mutagenesis is performed (STEP1) as described earlier (Sidhu, S. S. (2000) *Methods Enzymol.* 328:333-63). Typically a 10 microgram single-stranded template reaction will yield a library size of 10^8-10^9 transformants. This mutagenesis reaction is transformed into DH5-alpha cells and a maxiprep is performed on the CDRL3 library collection.

This collection of L3 Library DNA is transformed into CJ236 cells for preparation of L3 Library single-stranded template (STEP2). This creates the single-stranded template for incorporation of additional CDR mutagenesis. The mutagenized H3 library oligonucleotides (STEP 3) are annealed to the CDRL3 library template. For the template utilizing sequence shown in Fig. 18, there are ten lengths of CDRH3 used in the initial design: sizes 9-18. For each size of CDRH3, separate reactions are performed. Therefore in step 3, ten separate reactions are performed utilizing 20 micrograms of single-stranded template for each reaction. Within each size of CDRH3, nine walkthrough amino acids are utilized. Therefore nine degenerate oligonucleotides are pooled together for each reaction as for size 9 (VH_3_9*) and similar number for size 10 (VH_3_10*), etc. Each 20 microgram single-stranded template reaction yields more than 10^9 transformants. Therefore, for this library,

which contains ten CDR sizes, the total CDRL3/CDRH3 library is greater than 10^{10} total transformants. Each mutagenesis reaction is transformed into DH5-alpha cells, and after maxi-preparation of plasmid DNA, the DNA can be transformed into an appropriate expression-display cell line such as DH12S cells for APEx display and screening (as described in Harvey et al. PNAS 101 (25): 9193. (2004)) or the plasmid can be further transformed into CJ236 cells for further incorporation of mutations in other CDRs such as CDRH1, CDRH2, CDRL1 and CDRL2 (STEP 4).

EXAMPLE 1
METHODS FOR BIOINFORMATIC-GUIDED IDENTIFICATION OF UNIVERSAL ANTIBODY LIBRARY SEQUENCES

In this example, universal antibody library sequences are identified and selected using bioinformatics and the criteria of the invention.

Briefly, the Kabat electronic database containing expressed, *i.e.*, rearranged immunoglobulin sequences, was searched using certain filter algorithms. In particular, the filter algorithms were designed to identify only human sequences that were expressed in response to a particular antigen class. The antigen class selected was protein-based antigens/targets because this is a tractable set of targets for the development of human therapeutics. It is understood, however, that the database is just as easily queried for other antigen classes, *e.g.*, peptides, polysaccharides, polynucleotides, and small molecules as well as for antibody sequences derived from other species such as primate, mouse, rat, or chicken sequences, etc., for the development of, *e.g.*, therapeutics for veterinary application. The foregoing criteria were applied to an initial set of 5971 V_H sequences (it is noted, however, that this set of sequences can increase in number as additional sequences are cloned and entered into the database).

The above search and filter analysis returned a dataset of ~380 V_H gene sequences that represent non-redundant rearranged human antibody clones recognizing protein antigens. The next step involved the designation of the germline precursor that generated these rearranged gene sequences, followed by a frequency analysis of these candidate germline sequences. In other words, a determination as to whether there are optimal or high frequency germline framework sequences for protein antigens. In order to determine the germline sequences employed by the rearranged genes in the filtered V_H sequences (from Kabat), V BASE was used. V BASE is a comprehensive directory of all human germline variable region sequences compiled from over a thousand published sequences, including those in the current releases of the Genbank and EMBL data libraries (see respectively, *e.g.*, Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410 and

publicly accessible databases run by the Centre for Protein Engineering MRC Centre, Hills Rd, Cambridge, UK, CB2 2QH). Currently there are 51 functional V_H segments grouped into 7 families: (*i.e.*, V_H 1-7), 40 functional V_K segments grouped into 7 families: (*i.e.*, V_K I-VII), and 31 functional V_λ segments grouped into 10 families: (*i.e.*, V_λ 1-10). A batch BLAST of V BASE germline sequences against the filtered Kabat sequences (~380 V_H) was performed to identify V_H germline genes (and families) that most frequently contribute sequences that are expressed (rearranged). The analysis for example, identified that six of the eight most highly represented frameworks (frameworks 1, 2, and 3) belong to the V_H3 germline family, and members of the V_H1 and V_H4 families formed a group of intermediate representation. A frequency analysis was performed on the germline V_H frameworks (1,2, and 3) sequences that are represented in the filtered Kabat database (Figures 3 and 4). For this analysis up to 4 somatic mutations (dotted bars) were permitted during the classification of the rearranged sequences to a germline gene. Identical matches between the filtered Kabat sequences and the germline genes are identified with open bars. The threshold line is set to identify those germline genes that are most frequently represented in rearranged sequences of antibodies that recognize protein/peptide antigens (Figures 3 and 4).

The identification of highest frequency V_H3 frameworks has an important consequence. Selection of the V_H3 initial framework sequence dictates the corresponding CDR sequence diversity and size limitations as defined by standard canonical structures (see, e.g., Chothia, C., *et al.*, Structural repertoire of the human VH segments. J Mol Biol, 1992. 227(3): p. 799-817 and Tomlinson, I.M., *et al.*, The Structural repertoire of the human V kappa domain. Embo J, 1995. 14(18): p. 4628-38). The V BASE-filtered Kabat BLAST search also identifies positions that are so-called 'hotspots' for somatic hyper-mutation that can be mutated during affinity maturation of candidate molecules. The preliminary Kabat–V BASE results identified four highly utilized V_H3 frameworks; 3-07, 3-11, 3-23, and 3-33 and V_H1 frameworks; 1-e, and V_H4 frameworks; 4-34 and 4-30.4 (Fig. 5). In choosing multiple starting frameworks added structural diversity was created outside of the CDRs for potential antigen binding. Comparative analysis of all six J_H sequences (that encode framework 4) indicates that four of these sequences are identical and there exist only two amino acid differences in the other two sequences. This sequence conservation allows for the use of a common framework 4 for all seven framework families minimizing the generation of non-functional diversity (Fig. 5).

Thus, it was demonstrated that a manageable set of antibody acceptor sequences can be rationally identified using the criteria of the invention and using a bioinformatic approach and existing antibody databases. Moreover, the identification of these

sequences provides the foundation for maximizing intelligent CDR diversity within the universal antibody library, as discussed below.

EXAMPLE 2
METHODS FOR DESIGNING CDR DIVERSITY FOR UNIVERSAL ANTIBODY LIBRARIES

In this example, methods for optimizing the CDR diversity of a universal antibody library are presented.

The choice of candidate frameworks, as previously noted, dictates both the CDR sizes to be introduced and the initial amino acid sequence selection. All six chosen V_H3 gene families have the same canonical structures of 1-3. Canonical structure 1-3 requires CDR1 and CDR2 to have, respectively, 5 and 17 amino acid loops. A CDR amino acid frequency analysis of the Kabat and V BASE databases allows identification of the CDR amino acids for 1) absolute sequence conservation, 2) the first round of diversity generation, and 3) subsequent affinity maturation by mimicking somatic hypermutation. The design of each CDR in the heavy and light chain variable regions are discussed sequentially, below.

To design the first CDR of the heavy chain, hereafter "VH-CDR1", the above criteria are considered as follows: CDR positions that are conserved in both the germline and rearranged genes are fixed; CDR positions conserved in the germline but variant in rearranged genes are fixed in the initial library construction but allowed to be mutated during affinity maturation; and CDR positions that exhibit diversity in the germline and rearranged sequences are positions for incorporating diversity using mutagenesis, for example, walk-through mutagenesis (WTM™). Starting with the six identified V_H3 gene families (*i.e.*, 3-07, 3-21, 3-23, 3-30.5, 3-48, and 3-74), comparative V BASE analysis of the germline 5 amino acid CDR1 sequence reveals that S31, Y32 and M34 are conserved among the six genes (Fig. 6, left panel). A frequency analysis of all rearranged 5 amino acid CDR1 sequences in the filtered Kabat dataset (Fig. 6, right panel) illustrates three important findings: first, Y32 is highly conserved, second, the conserved germline S31 and M34 positions are subject to subsequent somatic mutations, and third, CDR1 positions 33 and 35 are neither conserved in the germline nor in rearranged antibody sequences.

Accordingly, in VH-CDR1, Y32 is fixed and never subject to any alteration as the strict conservation of Y32 indicates strong selective pressures for its preservation. CDR1 positions 33 and 35 are sites for creation of initial CDR1 sequence diversity by mutagenesis, *e.g.*, WTM™. Positions S31 and M34 are initially "fixed" but are identified as sites for mutagenesis during affinity maturation in any scFv candidate clones. The reason for not creating diversity at all sites is to restrict the initial diversity of the library to facilitate expression and display.

From the above Kabat frequency analysis, CDR1 has a "wild type" consensus sequence of SYAMH. The residues A33 and H35 are chosen as wildtype sequences due to their highest frequency in Fig. 6. In introducing subsequent amino acid diversity, the CDR1 sequence would then be SYXMX, where X denotes the position where mutagenesis, *e.g.*, WTM, is conducted. For example, when mutagenesis, *e.g.*, WTM, is conducted on the tyrosine residue in CDR1 positions 33 and 35, the desired resulting CDR1 sequences are SYYMH, SYAMY and SYYMY. In this instance, the effects of introducing an aromatic side chain are explored. The oligonucleotide codon sequence for the wild type A33 position is GCX. If replaced by Y33, the needed corresponding oligonucleotide sequence would be TAY. Thus for an A33 →Y33 oligonucleotide mix, the resulting codon sequences are (G/T)(A/C)C. The generated A33 →Y33 oligonucleotides in this case can also have codon permutations coding for glycine (GCC), aspartate (GAC) and serine (TCC). These additional "by-products" contribute to additional diversity at position 33. For the next WTM™ position 35, the wild type codon sequence for H35 would be CAY and if replaced with Y35, the oligonucleotide sequence required would be TAY. Thus for an H35→Y35 mix the resulting codon sequence is (C/T)AC. In this case, there would be no additional amino acid "by-products" being formed.

In another approach, byproducts are avoided by employing look-through mutagenesis (LTM) which typically requires the synthesis of an oligonucleotide for each desired change but eliminates any by-products (noise).

To design the second CDR of the heavy chain, hereafter "VH-CDR2", the above sequence analysis in the V BASE and Kabat databases was approached in a similar manner as for VH-CDR1 above. A frequency analysis was performed for VH-CDR2 sequences and an alignment of germline CDR2 sequences from the six candidate frameworks was constructed and a threshold frequency of 10% was selected (Fig. 8).

Starting with the same V_H3 gene families (3-07, 3-21, 3-23, 3-30.5, 3-48, and 3-74), V BASE (Fig. 8, left panel) and filtered Kabat (Fig. 8, right panel) frequency analysis shows that CDR2 positions I51, Y59, A60 and G65 are conserved in all germline and most rearranged genes and therefore must be invariant in a synthetic CDR2. The above Kabat frequency analysis indicates that VH-CDR2 would have a "wild type" consensus sequence of GISGGTTYYADSVKG. Because VH-CDR2 positions 54, 55, 58, 61, 62, 63, 64 display sequence conservation in the germline but are subject to subsequent somatic mutations, and are therefore "fixed" but allowed to be mutated during affinity maturation. For initial CDR2 diversity, investigational amino acids (underlined) are incorporated at positions 50, 52, 52a, 53, 56 and 57 (XIXXXGGXXYYADSVKG) and introduced by mutagenesis, *e.g.*, WTM™.

WTM, unlike random mutagenesis, allows predetermined placement of particular amino acids. For example, to perform WTM™ of CDR2 with a tyrosine (Y) residue at positions 50, 52, 53, 56 and 57 (underlined), the desired resulting WTM™ CDR2 sequences include the following (alterations are underlined): single substitutions (YIXXXGGXXYYADSVKG, XIYXXGGXXYYADSVKG and etc.), double substitutions (YIYXXGGXXYYADSVKG, YIXXYGGXXYYADSVKG and etc), triple substitutions (YIXXYGGYXYYADSVKG and etc.), quadruple substitutions (YIYYYGGXXYYADSVKG or YIXYYGGYXYYADSVKG) quintuple substitutions (YIXYYGGYYYYADSVKG), and sextuplet substitutions (YIYYYGGYYYYADSVKG). Typically, 2-3 substitutions per CDR are preferred and this can be readily achieved by oligonucleotide synthesis doping (see, *e.g.*, US20040033569A1 for technical details).

WTM™ for CDR2 using the nine pre-chosen WTM™ amino acids produces a library diversity of $9x2^6$ or 576 members. For comparative purposes, CDR2 saturation mutagenesis of six positions with all twenty amino acids would be 20^6 or 6.4×10^7. Accordingly, performing saturation mutagenesis on the 12 "non-fixed" positions of CDR2 alone, the library diversity would be 20^{12} or 4×10^{15} which is beyond the capabilities of current library display and screening technology. This illustrates an advantage of the invention which, by contrast, allows for a smaller but more representative library to be constructed. Indeed, the methods of the invention provide for, a manageable library in some CDR positions in order to identify the first generation of binding molecules. Subsequent affinity maturation mutagenesis in the other CDR positions then optimizes those identified binding molecules.

To design the V_H CDR3 diversity, CDR3 sequences of antibodies from the Kabat database were aligned according to their size and antigen class. Lengths of CDR3s of antibodies recognizing non-protein (in shaded bars) and protein/peptide antigens (open bars) are shown and fitted to trend lines (solid for the former and dotted for the latter) (Fig. 11). A frequency analysis of the 13 amino acid CDR3 sequences from the filtered Kabat dataset was also performed and a threshold frequency of 10% was selected (Fig. 11). Because CDR3 size and amino acid residue frequency analysis is performed using, *e.g.*, the immunoglobulin D and J gene rearranged sequences, there are no CDR3 germline equivalents for direct filtered Kabat and V BASE comparisons. Nonetheless, a filtered Kabat database was examined and search results indicated that, in terms of CDR3 loop size, there is a normal distribution curve ranging from 6 to 24 amino acids with a crest at approximately 13-16 amino acids (Figure 10).

Without a parallel VBASE-to-Kabat comparative analysis for CDR3 positions, a filtered Kabat frequency analysis (Figure 11) for each rearranged CDR3 size was performed. Within each size classification, enumerating the most frequent amino acid at

that CDR3 position results in a consensus "wild type" sequence. Surprisingly, this "consensus" approach identifies particular amino acids under high selective pressures. For example, in a 13 amino acid sized CDR3, position 101 was highly conserved as an aspartate (Figure 11). Therefore, as above in designing the diversity of VH-CDR1 and VH-CDR2, D101 is maintained as a "fixed" residue position in the synthetic 13 amino acid VH-CDR3. The VH-CDR3 positions 96, 98, 100c, and 102, however, show a higher preference for some amino acids and are therefore preliminarily "fixed" but then mutagenized during affinity maturation. The frequency distribution indicates that CDR3 positions 95, 97, 99, 100, 100a, 100b, 100d, and 100e did not show any preferential amino acids. Thus in the 13 amino acid CDR3 sequence, the formula XGXSXXXXYXXDY represents the positions (underlined) that are sites of diversity using, *e.g.*, mutagenesis such as WTM. A similar analysis can be conducted for all sizes of CDR3 sequences between 8 and 20 amino acids. Figure 10 illustrates that this size range encompasses a majority of length diversity found in CDR3 of antibodies that recognize proteinaceous targets/antigens.

EXAMPLE 3
METHODS FOR GENERATING POSITIONAL VARIABILITY PROFILES (VP) FOR ANTIBODY CDRs USING BIOINFORMATICS

In another approach, universal antibody libraries (UALs) were designed by determining the Positional Variability Profiles (VP) for CDRs expressed *in vivo*. The Positional Variability Profiles represent the cataloging of the different amino acids, and their respective rates of occurrence, present at a particular position in a dataset of aligned sequences of naturally expressed antibodies.

Therefore, determination of VP entails two steps, e.g., step 1): collection and selection of (a dataset of) *aligned* amino acid sequences that shares one or more defined properties of interest to create a dataset. Separately aligned CDR1, CDR2, and CDR3 sequences from either the V_H or V_L form the initial datasets for typical purposes. Several approaches in deriving CDR datasets with corresponding VP outcomes are available. Typically, for conducting step (2), (CDR) datasets are enumerated for amino acid variability and their relative frequencies for each aligned position (Figure 43). The VP for each CDR dataset then identifies the desired characteristics of a given CDR position for further introduction of diversity representation.

For conducting step 1, a database of aligned sequences is assembled. Sequences are selected that share one or more defined features of interest. For example, the proteins may have identifiable motifs, domains, and/or are evolutionarily related family members to permit whole or portioned sequence alignment between them. The starting input dataset can be derived from a prior compilation of previously characterized and

grouped sequences such as the Kabat database of endogenously expressed mature antibodies.

From the Kabat database, human immunoglobulin and, in particular,VH sequences were selectively collected for the starting base dataset.. Typically, the root germline origin for each rearranged human VH sequence is determined by comparative analysis. The corresponding germline foundation is termed the "originating subfamily" (STEP 1 in Figure 41). Additional CDRs in the VH sequences using the parameters set forth by Contact Definition within this starting "base dataset" can be identified and delineated. The designation of CDRs and their comprising amino acids can also be described by Kabat, Chothia or any other suitable definitions (STEP 2 in Figure 41).

Within the starting human VH sequence "base dataset", the compiled VH sequences are still likely to possess vastly different characteristics. The VH framework sequences will vary in regards: 1- family groupings (VH1, VH2, VH3, VH4 etc). 2- "originating subfamilies", 3- CDR lengths, 4- CDR canonical structure classes, 5- antigen specificity among others. Due to sequence disparity among the "base dataset" members, trying to derive a coherent analysis may require further selection from the starting "base dataset" such that datasets that share one or more of elected properties of interest, can be identified. Constituent members sharing those respective properties can produce a more "standardized" set of sequences for meaningful comparative analysis within the subgroup. This process can be iterated, resulting in the generation of smaller datasets of higher degree of relationships and such is exemplified in Figure 43.

CDRs can be classified as follows. Beginning with the non redundant "base dataset" of all human VH sequences, only sequences generating VH1 sequences were further selected (Figure 43). Non-redundancy filtering removes duplicate antibody sequence deposited against the same antigen. If there are different antibodies raised against the same antigen, these sequences are retained in the database. Within the VH1 sub-group, CDR1 and CDR2 sequences are identified and then further partitioned as CDR1 or CDR2 subgroups. It should be noted that in CDR partitioning within VH families, the CDR1 or CDR2 sub-groups can still be populated with CDRs of different lengths. VH1 CDR2 occurs in both 13 and also 15 amino acid lengths. For CDR1 and CDR2 lengths 6 and 13 amino acids are selected respectively and the generated datasets are named VH-1_CDR1_6 and VH-1_CDR2_13 respectively (Figure 43).

Canonical structures are classified as follows. Within the VH1 CDR2 sequences, another sub-group partitioning to further filter between those that were either of canonical structure 2 (CS2) or canonical structure (CS3), is performed. Canonical structures can be defined by distinguishing signature residues at key residues. The amino acid residues and position depend on which definition of a CDR is utilized. In this example, VH1 CDR2 sequences incorporating a V,A,L or T at amino acid position

71 denoted canonical structure 2 whereas, an R at the same amino acid position signified canonical structure 3. These operations generated datasets named, respectively, VH-1_CDR2_13_CS2 and VH-1_CDR2_13_CS3 (Figure 43).

The sub-grouping of VH-1_CDR2_13_CS2 and VH-1_CDR2_13_CS3 (Figure 49) reveals slightly different variability profiles between them and the more VH-1_CDR2_13 general collection (Figure 48). For example, in VH-1_CDR2_13_CS2 the "A" is not one of the more preferred amino acids at position 52a. For VH-1_CDR2_13_CS3, an "M" is found to be favorably introduced into position 51 which is not in either VH-1_CDR2_13_CS2 and VH-1_CDR2_13. A more dramatic example between VH-1_CDR2_13_CS2 and VH-1_CDR2_13_CS3 is that in the former, there is a near equal preference between A and T at position 57.

VH1 CDR1 has canonical class 1 (CS1) with the requirement of amino acid 6 residues generating the subgroup: VH-1_CDR1_6 (Figure 44). In this case, CDR1 CS1 distinguishing key amino acid signatures include for example; a T,A,V,G, or S at position 24, a G at amino acid position 26, and either a I,F,L,V, or S at position 29. Thus it is possible that the VH-1_CDR1_6 dataset can contain sequences not belonging to CS1 in that some 6 amino acid CDR variants do not have the requisite signature sequences. The main purpose of the universal antibody library is to best match framework sequences with CDR canonical structures and their variable sequences therein to obtain the most stable and functional configurations. Therefore, these non-CS1 sequences can contribute sequence "noise" in the dataset introducing amino acids not naturally optimized for CS1 stability and functionality. Thus, a further refinement can partition only those sequences having CS1 signature matches to generate the dataset VH-1_CDR1_6_CS1 (Figure 45).

Cross-CDR pair matching was performed as follows. The sub-groups above are examples of sequence collections that have been filtered and standardized in respect of VH germline sub-family, CDR length size and canonical structure. There are other parameters, outside the immediate CDR structural constraints, that can directly influence the endogenous *in vivo* CDR sequences. Indeed, the phenomenon that the CDR canonical structure can influence the CDR sequence within, can be demonstrated. It is possible that having one CDR canonical structure can influence both the canonical structure and sequence of another CDR. To demonstrate this CDR interdependency, CDR sequence analysis was performed when sub-groups where partitioned based on inter-CDR canonical structure pairings. For example, CDR sub-groups were collected whereby CDR1 (in this case just CS1) was grouped to either CDR2 CS2 or CS3 in the original antibody sequence composed thereof.

Following this rationale, VH-1_CDR1_6_CS1 was thus split into either VH-1_CDR1_6_CS1-2 (Figure 46) and VH-1_CDR1_6_CS1-3 (Figure 47) representing the

"pairing" of CDR1-CS1 with CDR2-CS2 and CDR2-CS3 respectively. Generally the variability profile is similar with the overall VH-1_CDR1_6 collection but there are individual CS preferences at particular CDR positions. At position 31 in VH-1_CDR1_6_CS1-2 (Figure 46), the "G" and "D" would not appear on the variability profile compared to either VH-1_CDR1_6CS1 and VH-1_CDR1_6_CS1-3. Between VH-1_CDR1_6_CS1-2 and VH-1_CDR1_6_CS1-3, position 33 also displays some variability outside. In the VH-1_CDR1_6_CS1-2 pairing, the "A" is the dominantly represented amino acid along with uniquely associated "G", "T", and "W" whereas; in the VH-1_CDR1_6_CS1-3 pairing the dominant amino acid is the "Y" with its' preferred variable amino acid of "D". Although not analyzed, VH-1_CDR2 also occurs with a canonical structure of "U" in a length of 15 amino acids that also can be selectively pair matched to their endogenous CDR1 sequences. It is predicted that the resulting VH-1_CDR1_6_CS1-U variability profile can be different from the above VH-1_CDR1_6_ CS1-3 and CS1-2 variability profiles.

In the converse, the interdependency between CDR2 sequences in relation to CDR1 canonical structure 1 was analyzed in similar fashion. Figure 50 illustrates the variability profile of VH-1_CDR2_13_CS2-1 and VH-1_CDR2_13_CS3-1. In this case though, the CDR2 variability profiles were nearly identical demonstrating that CDR2 design can function independent of CDR1 in this respect. However, for other frameworks such as VH4 family, the VH4 CDR2 has only one canonical structure (CS-1) with a CDR2 length of 12 amino acids. It is the VH4 CDR1 that involves three canonical structures CS-1, CS-2 and CS-3. In this case, unique CDR positional amino acid preferences from the resulting VH-4_CDR2_12_CS1-1, VH-4_CDR2_12_CS1-2, and VH-4_CDR2_12_CS1-1 variability profiles are anticipated.

These above results demonstrate that depending on both the CDR1 and CDR2 canonical structures chosen to be utilized as the acceptors, amino acids can be "fine-tuned" depending on which amino acids will be introduced in the various CDR amino acid positions to replicate the employed natural diversity, e.g., by matching sequences most likely to be found with other sequences if there is cross-CDR stabilization.

CDR antigen classification was performed as follows. Once grouped based on structural classifications, the collected members based on antigen specificity can be classified (Figure 43). There can be a correlation of preferred amino acids within the CDRs for a given antigen class and this was observed for antigen class preference for certain frameworks. Thus it is possible to add an additional parameter, antigen specificity, in the partitioning of CDR sequences.

Broadening CDR sequence collections was performed as follows. The above analysis has demonstrated the addition of screening combinations of multiple parameters to generate a sub-group of interest. This has the effect of "narrowing" the selected CDR

sequences for variability profiles. However, there can be the occasion to perform the reverse, that is, to obtain larger datasets with lower degree of homogeneity or shared properties. This in effect is accomplished by combining different datasets (Figure 43). In this example, the variability profile of all CDR2s of length 12 with a canonical structure 2 (CDR2_13_CS2) irrespective of what VH family those CDR2 may be attributed to, was enumerated. This effectively gives a broader survey of all amino acids that contribute to CDR2_13_CS2 diversity.

The selective process for choosing CDR1 and CDR2 sub-group dataset is described in Figure 43. The important advantage of our process is that many different "selection" paths are possible, and each of them generates a different dataset and hence a different variability profile (VP).

This is exemplified in the figures below where the variability profiles for CDR1 are compared between VH1, VH3 and VH4. Although somewhat similar, one important difference occurs in position 34. For VH1, an "I" can be fixed, for VH3, an "M" can be fixed, and for VH4, a "W" can be fixed for CDR design. Another difference would relate under the use of 50% - 80% frequency considerations. For VH4, the two most frequently found amino acids at position 35 are "S" and "H". Collectively, the aggregate percentage of those two amino acids would be greater than 80%. As such, position 35 can be characterized as "fixed" and both "S" and "H" as forced co-products can be introduced at that position. In contrast, for both VH1 and VH3, the two most frequently found amino acids at position 35, their frequency of occurrence do not the aggregate greater than 80% and would be characterized as "variable" and subjected to mutagenesis (e.g., WTM) at that position.

EXAMPLE 4
METHODS FOR GENETICALLY ENGINEERING A UNIVERSAL ANTIBODY LIBRARY

In this example, the steps for making and assembling a universal antibody library using genetic engineering techniques are described.

Briefly, the V_L and V_H fragments of the antibodies are cloned using standard molecular biology techniques. The oligonucleotides encoding the framework and CDRs of the variable regions are assembled by the polymerase chain reaction (PCR). These V_L and V_H fragments are then subsequently linked with a poly-Gly-Ser linker (typically GGGGSGGGGSGGGGS) to generate single chain antibodies (scFv). The full-length molecules are then amplified using flanking 5' and 3' primers containing restriction sites that facilitate cloning into the expression-display vector(s). The total diversity of the

libraries generated depends on the number of framework sequences and number of positions in the CDRs chosen for mutagenesis, *e.g.*, using WTM.

Typically, the average diversity of the V_H library, using 9 amino acids to conduct WTM™, is 3.5×10^6 (6 frameworks x 9 amino acids x (2^2 for CDR1 x 2^6 for CDR2 x 2^8 for the 13 amino acids CDR3)). The diversity of the V_H library is an upper limit and the diversity of the V_λ and V_κ libraries is significantly smaller, thereby limiting the combined diversity of the complete scFv library from 10^{10} to 10^{11} which is within the range of the transformation efficiencies of bacterial systems.

Accordingly, 90 oligonucleotides are synthesized to encompass the frameworks of the V_H, V_λ, and V_κ libraries in addition to 2 oligonucleotides that code for the linker region and 2 oligonucleotides that encode for His and Myc immunotags at, respectively, the N and C-termini. In addition, a subset of 30-60 degenerate oligonucleotides displaying the diversity in CDRs 1, 2, and 3 of each of the three libraries are synthesized (total 90-180). These oligonucleotides are assembled by the Single Overlap Extension (SOE) PCR method to generate the libraries that include the necessary V_H-V_λ and V_H-V_κ combinations. Random clones from each library are then chosen for sequence verification and assessment of library quality.

Regarding CDR diversity, LTM is used to explore small perturbations within the antibody CDR loops (*e.g.*, one change per loop). For further improvement, WTM, which allows for the incorporation of more than one substitution within a CDR, is subsequently used to exhaustively screen the chemical landscape of the CDR(s). Using WTM, the wildtype amino acid and the desired amino acid variants are explored in targeted CDR positions by manipulating oligonucleotide synthesis. A mixed pool of oligonucleotides is synthesized where a subset of the oligonucleotides code for the wildtype and another subset code for the targeted mutation in a specific position. In the WTM procedure, at each step of the synthesis, the growing oligonucleotide chain is extended by one of two bases. One base encodes for the wild-type codon, while the other base belongs to a codon for the desired mutation.

EXAMPLE 5
METHODS FOR THE EXPRESSION AND DISPLAY OF A UNIVERSAL ANTIBODY LIBRARY

In this example, methods for expressing and displaying a universal antibody library for screening against targets, are described.

Briefly, a bacterial expression and display system is used which has a demonstrated reliability for expressing scFv molecules from libraries. The scFv format consists of the functional antigen binding units (V_H and V_L regions) joined together by a

linker peptide (Fig. 14). Such libraries of the invention augment the diversity of the natural repertoire and once constructed can be repeatedly screened for other antigens.

The scFv library is transfected into the recipient bacterial hosts using standard techniques. The expressed fusion-scFv proteins are expressed at an outer surface location which permits binding of fluorescently labeled antigens. Candidate proteins are individually labeled by FITC (either directly or indirectly *via* a biotin-streptavidin linkage). Those members of the library expressing suitable scFv clones that efficiently bind the labeled antigens are then enriched for, using FACS. This population of cells is then re-grown and subjected to subsequent rounds of selection using increased levels of stringency to isolate a smaller subset of clones that recognize the target with higher specificity and affinity. The libraries are readily amenable to high-throughput formats, using, *e.g.*, FITC labeled anti Myc-tag antibodies and FACS analysis for quick identification and confirmation.

Candidate clones are then isolated and plasmid preparations are performed to obtain scFv sequence information. The approach allows for a hypothesis-driven rational replacement of codons necessary to determine and optimize amino acid functionality in the complementarity determining regions (CDRs) of the V_H and V_L regions of the antibody. Comparative sequence analysis and individual clone affinity/specificity profiles then determine which clones undergo affinity maturation (see Example 6).

EXAMPLE 6
METHODS FOR PERFORMING HIGH-THROUGHPUT AFFINITY MATURATION OF CANDIDATES FROM A UNIVERSAL ANTIBODY LIBRARY

In this example, the steps for identifying and improving a candidate antibody from a universal antibody library using affinity maturation is described.

Briefly, in order to validate the power of the universal antibody library and the ability to take a candidate antibody molecule and refine the binding properties of the molecule, a commercially available antibody was designated as a test antibody and mutagenized (using, *e.g.*, WTM™/LTM™ technology), expressed, displayed, and improved according to the methods of the invention.

Briefly, the test antibody was mutagenized in a scFv format and then expressed and displayed using yeast display, although any of the above-mentioned bacterial display systems can also be used. Kinetic selections of scFv yeast displayed libraries involve initial labeling of cells with biotinylated antigen followed by time dependent chase in the presence of large excess of un-biotinylated antigen. Clones with slower dissociation kinetics are identified by SA-PE labeling after the chase period and sorted using a high speed FACS sorter. The left panel of Fig. 29 shows the resultant dotplot of

the wildtype control and sorting gate, the dotplot showing the library and the number of clones in the sorting gate, and the dotplot of the clones isolated from the library post-sorting. In the right panel of Fig. 29, data from dissociation assays for two affinity matured clones as compared to the wildtype protein were fitted to a single exponential curve to determine the dissociation rate constants (k_{off}). Clones 1 and 2 exhibit 5.2- and 4.3-fold slower k_{off} rates than the parent molecule.

DNA sequence verification of randomly chosen clones indicates that the libraries are of high quality with respect to desired mutational diversity, unintended point mutations, deletions, and insertions. This efficiency contrasts with random/stochastic mutagenesis strategies where uncontrolled introduction of various bases produces higher levels of undesired base change effects leading to low expression or antibody functionality due to unfavorable amino acid usage and inadvertent stop codons.

Moreover, tabulated sequence data from a test antibody LTM analysis indicates productive diversity with very little noise. The bottom panel of Fig 29 shows the wild type sequence and 29 separate mutations that increase the affinity of the parent molecule by 1.5-fold or better which were uncovered in all six CDRs). Several of these changes were isolated multiple times, for example in CDR3, three separate S to K changes, and two S to Q changes were found. By contrast, shaded columns indicate the CDR positions where changes were never found to increase the affinity for the antigen. Subsequently, the combination of all the discovered LTM single mutations into one library facilitates the isolation of clones that exhibit improved avidity among these high affinity mutations.

EXAMPLE 7
METHODS OF SCREENING A UNIVERSAL ANTIBODY LIBRARY FOR IDENTIFYING A THERAPEUTIC ANTIBODY CANDIDATE FOR TREATING HUMAN DISEASE

In this example, methods for screening a universal antibody library of the invention for identifying a therapeutic candidate are described.

Briefly, a chronic and devastating renal disease that has been recalcitrant to previous therapies was chosen as a target for screening against the universal antibody library of the invention. In particular, Chronic Kidney Disease (CKD) is recognized as a major public health care issue in the U.S. with over 20 million afflicted individuals. A major hindrance in understanding nephrogenic processes is the lack of suitable reagents that recognize renal specific biomarkers that identify 1) the different cell types involved, and 2) the participating molecules that influence differentiation on these cells. Antibodies that recognize these renal markers would significantly augment the current pool of reagents needed to investigate kidney organogenesis and disease diagnosis.

To understand renal biology, six kidney specific human antibody candidates, 1) a Na-H exchanger (isoforms NHE3, NHE8) (14,15), an anion exchanger (isoforms SLC26A6, SLC 27A7), an adhesion molecule Ksp-cadherin (16), and lipocalin, were identified for screening against the universal antibody library.

Hematologists have long benefited from monoclonal antibody (Mabs) reagents recognizing "cluster of differentiation" (CD) cell surface markers. Hematopoiesis, the process that generates the lymphoid and myeloid lineages, has often shown many advantages as a model developmental system. Much of the reasons for its success reside in the ease in which hematopoietic cells can be identified, isolated and manipulated by Mabs.

To assay therapeutic candidates identified in the above screen, diagnostic disease biomarkers, *e.g.*, neutrophil-associated gelatinase- associated lipocalin (NGAL) a biomarker in the detection of early acute renal failure (ARF), can be used . In addition, a disease target for therapeutic treatment, for example, glomerulonephritis, can be monitored with α3 (IV) collagen protein. These proteins are biotinylated using existing protocols to facilitate FACS visualization using streptavidin-phycoerythrin (SA-PE) and then subjected to 3-5 rounds of selections to identify a first round of antigen binders from the universal antibody library. The initial antibody candidates are then sequenced and tested for affinity with purified soluble proteins using a BIAcore assay. The antibody candidates are then affinity matured, if desired, as described in Example 6.

Laboratory experiments are then performed to determine their functionality in recognizing the antigen target using art recognized techniques such as immunohistochemistry, immunoblot biomarker diagnostics, and *in vitro* and *in vivo* antibody blocking experiments.

EXAMPLE 8
METHODS OF BIOINFORMATIC-GUIDED IDENTIFICATION OF UNIVERSAL ANTIBODY LIBRARY SEQUENCES USING FILTERING AND CLUSTER ANALYSIS OF GENE SEQUENCES

In this example, methods for identifying universal antibody library sequences using database analysis, are described.

Briefly, VBASE and KABAT were selected as the main source of data for the purpose of designing universal antibody libraries with optimal structural and functional diversity. VBASE is a database containing the DNA and polypeptide sequences of all human germline segments, aligned and annotated according to the Kabat CDR definitions (Kabat *et al.*) with the numbering scheme based on Chothia (Chothia *et al.*). KABAT is the most comprehensive database of rearranged antibody sequences from disparate species. To improve antibody affinity by introducing diversity in all the CDRs of both the light and the heavy chains, the contact definition scheme for CDRs

(MacCallum *et al.*) was selected as an alternative scheme to the Kabat (Kabat *et al.*) and Chothia (Chothia *et al.*) guided approach. The contact definition approach allows for the introduction of significant structural diversity and improve binding without affecting the structure stability of the antibody. Chothia numbering, however, is used because it is the optimal scheme to be used with the contact definition approach. In Fig. 30, a comparison of the three most used CDR definitions is shown. According to our choice the number of amino acids for framework

Table 1 Distribution of amino acids along heavy and light chains according to the contact CDR definition of MacCallum et al.

	FR1	CDR1	FR2	CDR2	FR3	CDR3	FR4
VH	29	6-8	11	12-15	37	9-?	12
VK	29	7-13	9	10	33	8-?	11
VL	28	7-13	9	10	33/35	8-?	11

The VBASE analysis was performed for all the germline V segments (51 VH, 40 VK, and 31VL) have been downloaded and parsed according to the above definitions and stored locally in 3 different filed in a format described by FR1-CDR1-FR2-CDR2-FR3 where FR refers to framework region, *e.g.*, as shown in Fig. 31.

From these datasets individual data for each FR or CDR can be extracted. In particular, sequences were built as FR1xFR2xFR3 (called FR123) where "x" is used as place-holders for CDRs 1 and 2. The resulting datasets are stored in a convenient and compact forma such as FASTA (Fig. 32)

For each of the three germline families the sequences in FR123 were used to generate a distance matrix to analyze their relationship in the framework space. All the identical sequences were collapsed into one and all sequences with high similarity were clustered together. Each cluster is a representation of similar structural characteristics. Hierarchical clustering and the corresponding trees of FR123 sequences were computed using UPGMA method within the PHYLIP Package [PHYL].First a distance matrix was computed using PROTDIST [PHYL] and the simple Kimura's formula. Then NEIGHBOR [PYHL] was conducted with the UPMGA algorithm (see following figures).

SeqhuntII [Johnson *et al.*] was used to download the full datasets (see Table 2) of human VH, VK and VL sequences from the Kabat Database [KBTDB] in ASCII format and stored locally in three different files (rawdata). These files have then been parsed and each kabat entry has been stored onto a local DBMS A java package (com.bioreninc.kabatDB) containing classes to parse and analyze the datasets above has been developed. The package also provides a number of methods to convert and write different assemblies of the input sequences that allow for the isolation and analysis of specific regions. The design of the package is flexible and permits easy switching

between numbering systems and/or CDR definitions. Methods for identification of Canonical Classes of CDR1 and CDR2 have been implemented.

Table 2 Number of human sequences downloaded from the Kabat database

	Human (H)
VH	5971
VK	2374
VL	2012

The Kabat analysis filters were configured such that the original datasets were filtered in several sequential steps using a java package (com.bioreninc.unilib) and some external tools (PROTEIN SPECIFICITY (com.bioreninc.unilib)). To analyze a subset of the stored rearranged immunoglobulin sequences that recognize only protein antigens (called PA filter henceforth) an appropriate filter was selected. Sequences stored in the KabatDB that lacked antigen annotations were excluded (Table 3).

Table 3 Size of datasets after PA filter

	H	Prot. Ant. (PA)
VH	5971	758
VK	2373	454
VL	2012	217

Table 4 Size of datasets after CF123 filter

	H	PA	Redundancy (0.95)
VH	5971	903	547
V□	2373	618	268
V□	2012	310	140

To avoid bias caused by the redundancy of the database, some sequences were filtered out using CD-HIT and java tool (to double-check the results). The algorithm is based on the generation of clusters of sequences having above a chosen threshold (95%) of identity, followed by the selection of a representative sequence from each cluster. The similarity search was done on full-length sequences (*i.e.* FR1-CDR1-FR2-CDR2-FR3-CDR3-FR4).

As most of the library design relies on the fine-tuned analysis of framework regions 1, 2 and 3, it was very important to have complete sequence data to avoid misclassifications and/or wrong assumptions (Table 5).

Table 5 Size of datasets after 95% redundancy filter

	H	PA	Redundancy 0.95	CF123
VH	5971	903	547	378
VK	2373	618	268	169
VL	2012	310	140	78

The library design was split into two connected sub-projects: frameworks and CDRs. The frameworks selected from the human repertoire were germline framework segments 1, 2, and 3 representative of their usage in rearranged immunoglobulin sequences. For this purpose, the filtered datasets were first parsed and re-wrote in a FR123 format. This particular format was used during the entire framework selection process to determine how the germline framework usage was distributed on the input dataset, so that the most popular families would be identified. For this purpose a classification was executed for each sequence in the rearranged dataset and a similarity analysis was conducted using BLASTP [BLST] for each Kabat-FR123 dataset against the associated VBASE-FR123 and then parsed. The results were select to reveal hits with the highest similarity score. The cardinality of each family cluster was compared and the most popular ones were chosen as targets for the frameworks choice.

Table 6 Selected germline families

Chain	Selected Sub-Families
VH	VH-1 VH-3
VK	Vk-I Vk-III
VL	VL-1 VL-2 VL-3

Table 7 Germline family usage obtained from cluster analysis.

	sub-family	sequences	coverage
VH	VH-1	103	27%
	VH-3	153	41%
	VH-4	93	25%
VK	VK-I	80	47%
	VK-III	57	34%
VL	VL-1	22	28%
	VL-2	19	24%
	VL-3	33	42%

To visualize clusters from FR123-Kabat datasets, the PHYLIP Package [PHYL] was used (see Figs. 36, 37, and 38). Trees were obtained using distance methods (UPMGA). Distance matrices were computed using Kimura's formula. The resulting arrangement of the trees were matched with the previous blast analysis.

After selecting the germline families of interest, the analysis was fine tuned by investigation within the highly utilized framework segments and their relative canonical structure. Each VBASE dataset was blasted against the related selected Kabat dataset and parsed for output yielding only sequences with high similarity. Each VBASE framework segment was then ranked according to the number of computed high similarity hits. Finally, for the most popular VBASE clusters (within the selected families), representatives of the highest rated of its members, were chosen. If desired, the highest ranked sequence can be excluded for the most representative segment of the selected cluster (*i.e.* the segment which is at minimum distance from all the other segments within a cluster; see Table 8).

Table 8 Selected germline framework segments

V_H	CS	V_κ	CS	V_λ	CS
1-e	1-2	I-L1	2-1	1b	13-7
3-30*	1-3	III-A27	6-1	2a2*	14-7
3-23	1-3	III-L20	2-1	3l	11-7
3-07	1-3			3r	11-7
3-11	1-3				
4-30.4	3-1				
4-34	1-1				

EXAMPLE 9
METHODS FOR DESIGNING CDR DIVERSITY FOR UNIVERSAL ANTIBODY LIBRARIES USING EXTENDED CDR ANALYSIS

In this example, methods for designing CDRs for universal antibody libraries, are described.

Briefly, the selected frameworks (above) were used to guide the CDR selection and design: both lengths and sequences of the CDRs were specifically designed for each selected framework family to provide full compatibility and optimal diversity. The lengths of CDRs 1 and 2 were selected according to the canonical structure of the selected germline frameworks (VH-1, VH-3, VK-I, etc.). Starting from this basis a full analysis was performed for the subsequent design of CDR 1 and 2. The original Kabat dataset has been filtered only for completeness of the frameworks 1, 2 and 3 and for redundancy (95% similarity threshold) (see Table 9). For framework-CDR compatibility, no specificity filters were used.

Table 9 Filtering of original Kabat dataset for CDR design.

	Starting dataset	CF123	R95
VH	5971	2842	1865
VK	2373	859	471
VL	2012	1282	744

For each chain class (VH, VK, VL) the input sequences have been classified according to the selected germline families and binned in different datasets (see Table 9). These results have been obtained using the BLAST software [BLST] and parsing the results as described above.

Within each selected sub-family the length distribution of both CDR1 and CDR2 following the canonical structures classification was analyzed. More details about this analysis are discussed in the following section together with the adopted design strategy. A summary of the results is represented in Table 10.

Table 10 Classification of sequences for CDR design

	Sub-family	Sequences	Coverage
VH	VH-1	375	20%
	VH-3	761	41%
VK	VK-1	234	50%
	VK-III	136	29%
VL	VL-1	178	29%
	VL-2	185	25%
	VL-3	247	34%

The CDR 1 and 2 length was determined as follows. The VH germline family has CDR1 lengths 6 and 8, the last one being present only in VH-2 that is not use. CDR2 length varies form 12 up to 15, 13 being the most common and the one required by the selected frameworks.

The germline VH-1 always has CDR1 with 6 amino acids and CDR2 with 13 amino acids (canonical structures 1-3, 1-2). In the rearranged dataset ~97% of the sequences identified as VH-1 had CDRs of these lengths. Typical framework criteria are: 1-e CDR1 length: 6 CDR2 length: 13 with an expected class coverage of >97%

The germline VH-3 always has CDR1 with 6 amino acids and CDR2 with 13 and 15 amino acids. The frameworks selected have CDR lengths 6 and 13 respectively, so length 15 was not used for CDR2. Here the data showed that in 99% of the rearranged sequences CDR1 has length 6 as expected; in 81% CDR2 has length 13. The remaining usage space of CDR2 is most probably covered by canonical structures 1-U and 1-4 that have length 15 (in particular 3-15 has some usage popularity). Typical framework criteria are: 3-07, 3-11, 3-23, 3-30* CDR1 length: 6 CDR2 length: 13 with an expected class coverage of: ~81%.

The VK germline family CDR1 has a number of amino acids varying from 7 to 13, the most popular having 7 and 8 amino acids. The CDR2 always has 10 amino acids.

The germline VK-I always has CDR1 with 7 and CDR2 with 10 amino acids. The usage in the rearranged sequences shows a perfect match with germline information. Typical framework criteria are: I-L1 CDR1 length: 7 CDR2 length: 10 with an expected class coverage of: >97%.

The germline VK-III has CDR1 with 7 and 8 amino acids and always CDR2 with 10 amino acids. Here the data show that 50% of the sequences have CDR1 length 7 and ~48% have length 8. Such results were obtained because of the presence of two different and very common canonical structures. The CDR2 length is 10 in >98% of the sequences. With the selected frameworks the lengths of CDR1 are provided so that the

expected coverage was 98% of the usage space for the subfamily Typical framework criteria are: III-A27, III-L6 CDR1 length: 7 and 8 ; CDR2 length: 10; and with an expected class coverage of: >98%.

The VL germline family CDR1 has a number of amino acids varying between 7 and 10, where 8 is not common and so selectively excluded. Lengths 7, 9 and 10 are all quite frequent in the family.

The germline VL-1 has CDR1 with 9 and 10 amino acids and CDR2 with 10 amino acids. Data for rearranged sequences show that ~74% have CDR1 of length 9 and ~99% of them have CDR2 of length 10. The length 10 for CDR1 was excluded for better fit and typical frameworks selected were: 1b CDR1 length: 9 CDR2 length: 10 with an expected class coverage of ~75%.

The germline VBASE: VL-2 has both CDR1 and CDR2 with 10 amino acids. Typical framework criteria are: 2a2 CDR1 length: 10 CDR2 length: 10 with an expected class coverage of ~95%.

The germline VL-3 always has CDR1 with 9 and CDR2 with 10 amino acids. The selected 2 frameworks from this sub-family were chosen to provide more structural coverage because this sub-family is the most used in VL. Typical framework criteria are: 3r, 3l CDR1 length: 9 CDR2 length: 10; and with an expected class coverage of ~99%.

Table 11 Lengths of CDRs 1 and 2 distributed along the selected frameworks with selected sequences indicated

	CDR	length	sequences / total	
VH-1	1	6	369/375	√
	2	13	366/375	√
VH-3	1	6	752/761	√
	2	13	618/761	√
VK-I	1	7	228/234	√
	1	8	0/234	x
	2	10	232/234	√
VK-III	1	7	68/136	√
	1	8	65/136	√
	2	10	134/136	√
VL-I	1	7	0/178	x
	1	9	131/178	√
	2	10	45/178	x
	2	10	177/178	√
VL-2	1	7	1/185	x
	1	9	7/185	x
	2	10	176/185	√
	2	10	185/185	√
VL-3	1	7	244/247	x
	1	9	0/247	x
	1	10	0/247	√
	2	10	247/247	√

For each of the 15 selected CDRs (Table 11) a separate frequency analysis was executed to determine positional amino acid usage in the context of the selected framework. The main purpose was to provide a classification for each position within each CDR 1 and 2 into 2 different categories: a fixed position showing 1 or 2 dominant amino acids and positions for initial structural diversity, *i.e.*, mutagenesis.

A simple frequency analysis using EMBOSS/prophecy [EMB] was executed generating a matrix representing the positional amino acid usage. The output matrix has then been parsed and filtered in order to have relative frequency data for each position. The parser provides a very simple filter based on two thresholds (low and high). For each position the parser processes only amino acids with relative frequency above the low threshold until the cumulative frequency reaches the high threshold. If the high threshold is not reached, then the parser evaluates also the amino acids with relative frequency below the low threshold. A good low-high threshold combination was 10-80 because it provides good sensitivity for position classification. The parser output is visualized as frequency charts and the results are shown in the following figures.

Quantitative CDR classification.

Positions are classified as fixed when one or two amino acids are evidently dominant on the others. Usually in these situations the parser, with a good parameter tuning, is capable of filtering out the uncommon amino acids. The dominant amino

acid(s) are used as wild type in the CDR sequence; if two amino acids are dominant a "degenerate" wild type is used, which means that a mixed codon is synthesized to provide both the amino acids. The parameters chosen are very sensitive to identify high variability positions (WTM positions). In these positions there are no evident dominant amino acids but many different at med-low frequency. Here, the diversity can be represented using mutagenesis, *e.g.*, LTM or WTM with the most frequent amino acid as wild type.

In following figures, all the frequency charts and amino acid sequences for CDRs 1 and 2 that were developed for the universal library, are reported. The nomenclature of CDRs 1 and 2 is built as follow: CHAINTYPE-GERMLINEFAMILY_CDRTYPE-CDRLENGTH.. For example the name VH-1_CDR1-6 refers to Heavy Chain, family VH-1, CDR1 having 6 amino acids. The nomenclature of CDRs 3 is similar: it does not contain the germline family classification.

CDR design

The CDR3 of both heavy and light chains is the most variable region both in length and in sequence, providing most of the structural diversity of the antibody binding site. So, for each chain type, a length analysis both on the full dataset and on protein-specific chains was executed. A meaningful difference in the length distribution of the two datasets was found showing that protein antigens seem to prefer a slightly longer CDR3. The VH CDR3 has a distribution quite wide, so a lengths from 9 to 18 (~75% of the usage) were selected (see Fig. 10). The Vkappa CDR3 has a very narrow distribution and the most used lengths are 8 and 9 (see Fig 20). The Vlambda CDR3 has a slightly wider distribution and in this case, lengths 8, 9, 10 and 11 (see Fig. 24) were selected for the library.

On each selected lengths a frequency analysis was executed similar to the one described in CDR1 and CDR2 design. For all the lengths this analysis showed a high diversity in the locations in the middle of the CDRs and a few conserved positions close to the borders with framework regions.

As for CDR1 and CDR2, diversity regions in CDR3 were selected as high variability positions as targets for mutagenesis, *e.g.*, WTM. Residues chosen as wild type were the most frequent amino acids at each position. In some WTM positions, the wild type amino acid was chosen and typically, the presence of Gly was determined to be desirable. The WTM strategy was designed in a modified fashion: instead of choosing the "minimum-distance" combination of bases to provide the target and the wild type amino acids, mixed codons were designed in order to provide target, wild type and Gly amino acids (*i.e.* Gly is a required side-product). The following figures show all the frequency charts and the sequences of the CDRs chosen for the universal

antibody library. The nomenclature of CDRs 1 and 2 was built as follow: CHAINTYPE-GERMLINEFAMILY_CDRTYPE-CDRLENGTH. For example the name VH-1_CDR1-6 refers to the 6 residue positions of CDR1 of Heavy Chain, family VH-1. The nomenclature of CDRs 3 is similar but does not contain the germline family classification.

Additional CDR3 design conditions are as follows. Glycines are a necessity in CDR3 for functional loop structures. They are found in CDR3 in approximately 10-20% throughout the V_H CDR3 loop. Therefore, CDR3 regions were designed to accommodate multiple glycines throughout the loop. Therefore, in addition to the wild-type amino acid, glycines were required co-products in multiple V_H CDR3 positions. In position 95, an Asp was very common in the frequency table for antibodies against proteins and peptides, therefore, an Asp was used as the wild-type amino acid and Glycine as a required co-product (D/G) for WTM. Similarly for position 96, Arg was quite frequent, and therefore Arg was used as the wild-type amino acid and Gly as a required co-product (R/G). For positions 97-99, a Ser wass used and Gly as the base (S/G), since serine was a fairly common amino acid in CDR loops and is therefore well-tolerated. At position 101 Asp was used (held constant), and the position directly N-terminal of the Asp as well (Phe (D-1)).

For VH CDR3 lengths 10 and above, in the position that is two residues N-terminal to the Asp (D-2) (e.g. position 100 in VH_CDR3-10), a Tyr was used as the base amino acid. Tyr is also well-tolerated in the CDR loops of antibodies. The preponderance of Tyr N-terminal to the Asp increases with CDR3 loop length. Therefore, additional Tyr were added as the base amino acid as shown in Table 12. The remaining positions N-terminal of Asp101 until position 99 use a Ser as wild-type and Gly as a required co-product as shown in Table 12. As walk-through mutagenesis is performed, each CDR3 loop can be structured to create functional antibodies, since glycines are present for loop structure (generally 10-25%), and well-tolerated amino acids are present in the loop. Further functional binding interactions are gained through the walk-through amino acids and functional co-products.

A summary of identified CDR sequences for use in the universal antibody library of the invention is set forth below in Table 12. The names of the CDRs are standardized: the first field in the name is the germline family, the second field is the CDR type and the third field is the length of the CDR (example: VH1_CDR1-6 is the CDR1 of VH1 germline family having length 6). Single-letter positions are fixed positions; two-letters positions are combination positions where the synthesis is performed with a mix in order to have only 2 targeted amino acids (example: T-S); and two-letters at positions where the first is 'X' are WTM positions. The amino acid following the X is the wild type (example: X-V). Three-letters positions where the first

is 'X' are "coproduct-optimized" WTM positions. The amino acid letter following the 'X' is the wild type. The last amino acid (the one after the '/') is a required co-product.

Table 12 Summary of CDR Sequences for the Universal Antibody Library

CDR1

VH1_CDR1-6	30	31	32	33	34	35							
	T-S	S	Y	X-A	I-M	X-S							
VH3_CDR1-6	S	S	Y	X-A	M	X-S							
VK1_CDR1-7	30	31	32	33	34	35	36						
	S	S-N	X-Y	L	A-N	W	Y						
VK3_CDR1-7	S	S-N	N-Y	L	A	W	Y						
VK3_CDR1-8	30	30A	31	32	33	34	35	36					
	S	S-N	X-S	Y	L	A	W	Y					
VL1_CDR1-9	30	30A	30B	31	32	33	34	35	36				
	I	G	X-S	N	X-T	V	X-N	W	Y				
VL2_CDR1-10	30	30A	30B	30C	31	32	33	34	35	36			
	V	G	X-G	Y	N	Y	V	S	W	Y			
VL3_CDR1-7	30	31	32	33	34	35	36						
	X-S	K-Q	X-Y	A-V	X-H	W	Y						

CDR2

VH1_CDR2-13	47	48	49	50	51	52	52A	53	54	55	56	57	58
	W	M	G	X-G	I	X-N	P	X-I	X-S	G	X-T	T-A	N
VH3_CDR2-13	W	V	S-A	X-V	I	S	X-G	D-S	G	G-S	X-S	T-K	Y
VK1_CDR2-10	46	47	48	49	50	51	52	53	54	55			
	L	L	I	Y	X-A	A	S	X-S	L	Q-E			
VK3_CDR2-10	L	L	I	Y	G-D	A	S	X-S	R	A			
VL1_CDR2-10	L	L	I	Y	X-S	N	N-S	X-N	R	P			
VL2_CDR2-10	L	M-I	I	Y	E-D	V	S-T	X-N	R	P			

VL3_CDR2-10	L	V	I	Y	X-G	D	N-S	X-D	R	P

CDR3

VK_CDR3-8	89	90	91	92	93	94	95	96
	Q	Q	Y	X-N	X-S	X-T	P	X-L

VK_CDR3-9	89	90	91	92	93	94	95	95a	96
	Q	Q	Y	X-N	X-S	X-T	P	P	X-L

VL_CDR3-8	89	90	91	92	93	94	95	96
	Q	S-A	W	D	X-S	S	X-N	X-V

VL_CDR3-9	89	90	91	92	93	94	95	95a	96
	Q	S-A	Y	D-A	X-S	S	X-N	X-T	X-V

VL_CDR3-10	89	90	91	92	93	94	95	95a	95b	96
	Q	S-A	W	D	X-S	S	L-S	X-N	X-G	X-V

VL_CDR3-11	89	90	91	92	93	94	95	95a	95b	95c	96
	Q	S-A	W	D	X-S	S	L-S	X-N	X-G	X-P	X-V

VH_CDR3-9	93	94	95	96	97	98	99	100	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	F	D

VH_CDR3-10	93	94	95	96	97	98	99	100	100a	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-Y	F	D

VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	F	D

VH_CDR3-12	93	94	95	96	97	98	99	100	100a	100b	100c	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	F	D

VH_CDR3-13	93	94	95	96	97	98	99	100	100a	100b	100c	100d	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	F	D

VH_CDR3-14	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	F	D

VH_CDR3-15	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	F	D

VH_CDR3-16	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	X-Y	F	D

VH_CDR3-17	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	X-Y	F	D

VH_CDR3-18	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	100i	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y	X-Y	X-Y	X-Y	X-Y	F	D

Additional CDR3 designs incorporate greater diversity within several CDR3 positions, especially in the C-terminal region. This greater diversity more closely reflects the diversity observed in the Kabat database. This is reflected in CDR designs 2 and 3 (see tables below).

An alternate CDR3 design incorporates a tyrosine-rich design that incorporates tyrosines more broadly throughout the CDR3 loop. Although tyrosines are found broadly throughout the CDR3 loop, mixing a glycine codon with a tyrosine codon results in cysteine codons as well as the amber stop codon. The amber stop codon and broad cysteine incorporation would lead to non-productive antibody sequences. Therefore, glycines are included as coproducts at key positions where cysteines are observed in the Kabat frequency tables. Cysteines can form disulfide bridges to stabilize long CDR3 loops, therefore inclusion of cysteines at these key positions can be useful for CDR3 functionality. Increasing sizes of CDR3 also includes greater complexity in internal positions, and this is incorporated in the design principle.

Ideally tyrosines and glycines can be incorporated at all positions. In order to introduce these residues at every position without producing unwanted co-products such as the amber stop codon, an alternate oligonucleotide synthesis procedure is utilized where pools of codons are synthesized separately then combined and split for the following round of synthesis (E A Peters, P J Schatz, S S Johnson, and W J Dower, J Bacteriol. 1994 July; 176(14): 4296–4305.). In this process, two pools are utilized: the first pool utilizes the codon TMC, encoding Y and S, and the second pool utilizes the codon VRC, encoding H,S,R,N, and D. These pools therefore allow a hydrophobic contribution by tyrosine, and multiple polar contributions with the second pool. All diversity positions that are noted in green below are generated using split pools of these codons.

All these multiple CDR designs give multiple sublibraries of the universal libraries. Each design is tested empirically for overall fitness and performance against multiple antigens.

Table 13. Design 2

VH_CDR3-9	93	94	95	96	97	98	99	100	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	F	D

VH_CDR3-10	93	94	95	96	97	98	99	100	100a	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	F	D

VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-12	93	94	95	96	97	98	99	100	100a	100b	100c	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-13	93	94	95	96	97	98	99	100	100a	100b	100c	100d	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-14	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-15	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-16	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-17	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

VH_CDR3-18	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	100i	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	X-Y/A/N	X-Y/A/N	X-Y/A/N	F	D

Table 14. Design 3

VH_CDR3-9	93	94	95	96	97	98	99	100	101									
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	F	D									
VH_CDR3-10	93	94	95	96	97	98	99	100	100a	101								
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-Y/A	F	D								
VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101							
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	F	D							
VH_CDR3-12	93	94	95	96	97	98	99	100	100a	100b	100c	101						
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	F	D						
VH_CDR3-13	93	94	95	96	97	98	99	100	100a	100b	100c	100d	101					
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	F	D					
VH_CDR3-14	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	101				
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	F	D				
VH_CDR3-15	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	101			
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	F	D			
VH_CDR3-16	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	101		
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	X-Y/A	F	D		
VH_CDR3-17	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	101	
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	X-Y/A	F	D		
VH_CDR3-18	93	94	95	96	97	98	99	100	100a	100b	100c	100d	100e	100f	100g	100h	100i	101
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	X-Y/A	X-Y/A	X-Y/A	F	D	

EXAMPLE 10
OLIGONUCLEOTIDE DESIGN FOR INTRODUCING CDR DIVERSITY USING WTM AND EXTENDED WTM

Oligo construction can be carried out using the sequences set forth in Table 12.

Walkthrough and extended walkthrough (for CDRH3) were performed at the appropriate positions shaded in Table 12, where noted in the sequence denoted with an X. The X refers to the walkthrough amino acid, and the amino acid(s) following the (dash) – refer to the base amino acid and any required co-products denoted after a (slash) /. Positions in white with multiple amino acids listed denote an equal mix of those amino acids with the minimum number of co-products. This mixture reflects the predominant mixture of these amino acids present in variability profile.

For example, VH1_CDR1-6 is described as:

Table 15.

VH1_CDR1-6	30	31	32	33	34	35
	T-S	S	Y	X-A	I-M	X-S

If the walkthrough amino acid is chosen to be alanine, then the following codons are used for the above design:

Table 16.

5'-

30			31			32			33			34			35		
W	C	C	T	C	C	T	A	C	G	C	C	A	T	S	K	C	C

-3'

For position 30 (Chothia numbering), TCC encodes serine and ACC encodes threonine, therefore the most efficient mixture is WCC.

For position 31 TCC encodes S.

For position 32 TAC encodes Y.

For position 33 the walkthrough amino acid is identical to the base amino acid, therefore the base amino acid codon of GCC is used, encoding A.

For position 34 ATS encodes I and M, where ATC encodes I and ATG encodes M.

For position 35, the standard walkthrough procedure is used. TCC is serine, and GCC is the nearest alanine match. Therefore both G and T are required in the first position, C is required in the second position, and C is required in the third position. Therefore KCC is used, encoding A and S.

In practice, the oligonucleotides are synthesized with flanking regions complementary to the variable region of the antibody. Therefore, the following sequence is used:

Table 17.

VH1_1_6_WA	5'-	GCTTCCGGTGGCACATTC	30			31			32			33			34			35			TGGGTTAGACAGGCACCT	-3'
			W	C	C	T	C	C	T	A	C	G	C	C	A	T	S	K	C	C		

All 20 amino acids and unnatural amino acids utilizing the amber codon can potentially be walked through at the appropriate blue/green shaded positions. To exemplify, nine walkthrough amino acids are shown below.

Table 18.

			30			31			32			33			34			35				
VH1_1_6_WA	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	G	C	C	A	T	S	K	C	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WD	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	G	M	C	A	T	S	K	M	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WS	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	K	C	C	A	T	S	T	C	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WI	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	R	Y	C	A	T	S	A	K	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WP	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	S	C	C	A	T	S	Y	C	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WR	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	S	S	C	A	T	S	M	G	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WY	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	K	M	C	A	T	S	T	M	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WH	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	S	M	C	A	T	S	Y	M	C	TGGGTTAGACAGGCACCT	-3'
VH1_1_6_WN	5'-	GCTTCCGGTGGCACATTC	W	C	C	T	C	C	T	A	C	R	M	C	A	T	S	A	R	C	TGGGTTAGACAGGCACCT	-3'

To understand the nomenclature, VH1 is the framework VH1_1 refers to VH1 CDR1, VH1_1_6 refers to CDR size 6, and W refers to walkthrough and the final letter is the walkthrough amino acid. The above sequences exemplify walkthrough with A (alanine), D (aspartate), S (serine), I (isoleucine), P (proline), R (arginine), Y (tyrosine), H (histidine), and N(asparagine).

Oligo Construction using Table 12, was carried out using extended walkthrough and doping as follows.

Walkthrough and extended walkthrough (for CDRH3) were performed at the appropriate positions shaded in blue or green in Table 12, where noted in the sequence denoted with an X. The X refers to the walkthrough amino acid, and the amino acid(s) following the (dash) – refer to the base amino acid and any required co-products denoted after a (slash) /. Positions in white with multiple amino acids listed denote an equal mix of those amino acids with the minimum number of co-products. This mixture reflects the predominant mixture of these amino acids present in variability profile.

Table 19.

VH_CDR3-10	93	94	95	96	97	98	99	100	100a	101
	A	R	X-D/G	X-R/G	X-S/G	X-S/G	X-S/G	X-Y	F	D

This second example is given to exemplify the use of extended walkthrough mutagenesis with required co-products. The design in Table 12 for VH-CDR3 size 10 is shown above. The synthesized oligonucleotides for the alanine walkthrough is as follows:

For position 95, the base amino acid is aspartate, GAC. Alanine is GCC, and glycine is the required co-product GGC. Therefore G is in the first position, A, G, and C are in the second position and C is in the third position.

For position 96, the base amino acid is arginine CGC. Alanine is GCC, and glycine is the required co-product GGC. Therefore the first nucleotides of this position are C or G, the second nucleotides are G or C, and the third nucleotide contains a C.

For position 97, the base amino acid is serine and can be coded as TCC or AGC. Alanine is walked through with GCC, and glycine is encoded as GGC. For serine AGC is chosen because TCC combined with GGC produces a cysteine co-product (TGC), which is not generally desired in CDRs, since unwanted disulfide bond formation can occur. Therefore the AGC codon is chosen. Therefore the first nucleotide position contains A or G, the second position contains C or G, and the third coding position contains a C.

Positions 98 and 99 are identical to position 97, since they utilize the same base and required co-product amino acids.

Position 100 utilizes a tyrosine as a base amino acid TAC, and alanine is GCC. Therefore, the first coding position contains a T and G mixture, the second coding position contains A and C, and the third coding position contains a C.

These results are summarized below.

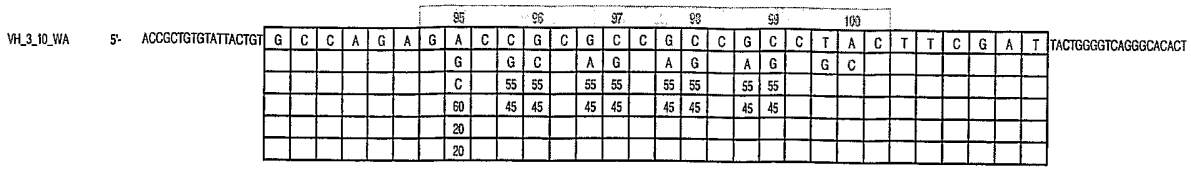
Table 20.

	95			96			97			98			99			100			
5'-	G	A	C	C	G	C	G	C	C	G	C	C	G	C	C	T	A	C	-3'
		G		G	C		A	G		A	G		A	G		G	C		
		C																	

In the preferred usage, flanking regions are added to the 5′ and 3′ regions to facilitate incorporation into the antibody sequence. In addition, since glycines represent 15-25% of the amino acid composition of CDRH3, doping can be performed achieve this approximate level of glycine incorporation.

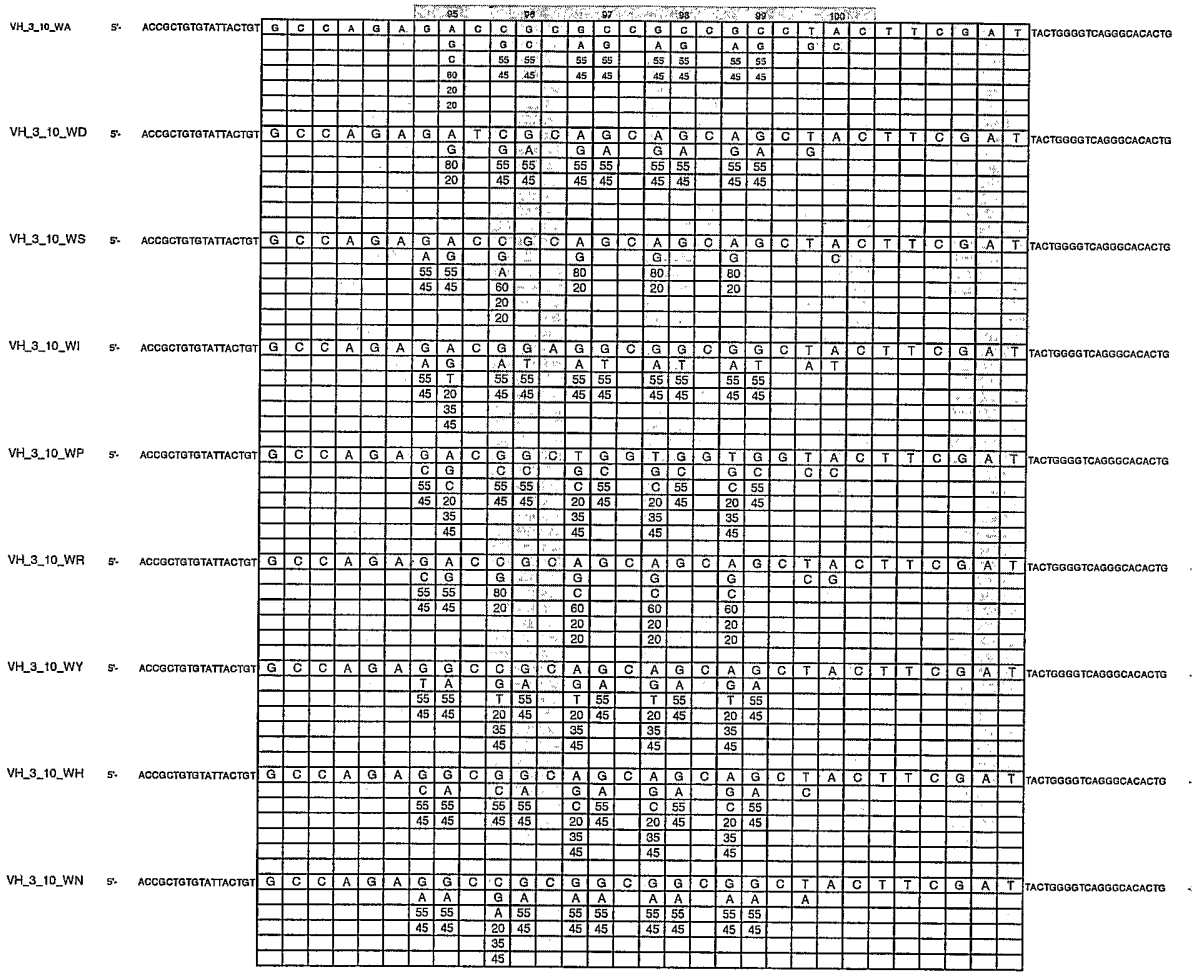
As an example, in position 95, the usage of glycine is defined by the percentage of G utilized in the second coding position. Therefore, to achieve 20% glycine incorporation, the percentage of G in the mixture was 20%. Similarly, in positions 96-99, the level of glycine incorporation was tuned to achieve an approximately 25% level of glycine incorporation while decreasing the level of co-product incorporation.

Table 21.



All 20 amino acids and unnatural amino acids utilizing the amber codon can potentially be walked through at the appropriate blue/green shaded positions. To exemplify, nine walkthrough amino acids are shown below for the size 10 VH CDR3.

Table 22.



Oligo Construction using CDRH3 Designs 2, 3, and 4 that utilize extended walkthrough and doping was performed as follows.

Walkthrough and extended walkthrough (for CDRH3) were performed at the appropriate positions shaded in blue or green in Table 12, where noted in the sequence denoted with an X. The X refers to the walkthrough amino acid, and the amino acid(s)

following the (dash) – refer to the base amino acid and any required co-products denoted after a (slash) /.

Table 23. Design 2 length 11

VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101
	A	R	X-S/G	X-S/G	X-S/G	X-S/G	X-S/G	X-Y/A/N	X-Y/A/N	F	D

This third example is given to further exemplify the walkthrough procedure, in this case with the additional CDR H3 designs.

All 20 amino acids and unnatural amino acids utilizing the amber codon can potentially be walked through at the appropriate green shaded positions. In this example histidine is walked through using design 2, proline is walked through using design 3, and serine is walked through using design 4. The sequence listing has these amino acids walked through the designs from length 9 to 18. Shown here is the analysis for length 11.

For design 2, histidine is the exemplary walkthrough/extended walkthrough amino acid. In position 95, the base amino acid is serine, and AGC is chosen over TCC to avoid the formation of cysteines when glycine is added as a co-product. Glycine is the required co-product and is GGC. Histidine is coded as CAC. Therefore A, G and C are in the first position, G and A are in the second position, and C is in the third position. The concentrations of the mixtures were doped to produce approximately 15-25% glycines, and favoring serine, the base amino acid.

Positions 96, 97, 98, and 99 utilize the same design as position 95.

Position 100 utilizes tyrosine as the base amino acid TAC, and alanine GCC, and asparagine AAC is required as a co-product. Histidine (CAC) is the extended walkthrough amino acid. Therefore the first position contains T, G, A and C, the second position contains A and C, and the final position contains C. Doping is performed to favor the base amino acid tyrosine.

Position 100a utilizes the same design principle as position 100.

After adding flanking regions to the primer, the oligonucleotide design for histidine walked through CDRH3 length 11 design 2 is shown below and in the sequence listing.

Table 24.

VH_3_11_WH Design 2 5'- ACCGCTGTGTATTACTGT GCCAGA AGCAGCAGCAGCAGCAGCTACAC TACAC TACTGGGGTCAGGGCACACTG -3'

95	96	97	98	99	100	100a
A	A	A	A	A	T	T
G	G	G	G	G	A	A
C	C	C	C	C	C	C
55	55	55	55	55	G	G
45	45	45	45	45	A	A
30	25	30	30	30	40	40
25	30	25	25	25	20	20
					20	20
					20	20

Table 25. Design 3 length 11

VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101
	A	R	X-D/G	X-D/G	X-S/G	X-S/G	X-S/G	X-Y/A	X-Y/A	F	D

For Design 3, proline is used for this example as the walkthrough/extended walkthrough amino acid.

In position 95, Asp (GAC) is the base amino acid, and Gly (GGC) is the required co-product. The walkthrough of proline (CCC) results in G and C in the first position, A, G, and C in the second position, and C in the third position. The glycines are doped to achieve between 15-25% frequency as in previous examples.

Position 96 utilizes the same design as position 95.

Position 97 utilizes serine TCG for proline walkthrough, since the walkthrough amino acid does not require a C or T in the third position. For a walkthrough amino acid requiring C or T, AGC can be utilized for the serine codon to avoid cysteine co-products. TCG is preferred over AGC because of the beneficial co-product of tryptophan TGG versus coding of AGG (arginine). Arginine is desirable, but CGC is already coded in the final mixture, making arginine redundant.

Therefore, TCG is used for Serine, GGG for the required glycine co-product, and CCG for proline. Therefore, T,G, and C are used in the first position, G and C in the second position, and G in the final position.

Position 98 and 99 utilize the same design as position 97.

Position 100 and 100a utilize the same design that uses tyrosine as the base amino acid (TAC), the required co-product Ala (GCC), and the extended walkthrough amino acid Pro (CCC). Therefore the first position contains T, C and G, the second position contains A and C, and the third position contains a C. Doping is performed to favor the base amino acid tyrosine.

With the flanking regions added, the oligonucleotide for the design 3 with proline walkthrough is shown below:

Table 26.

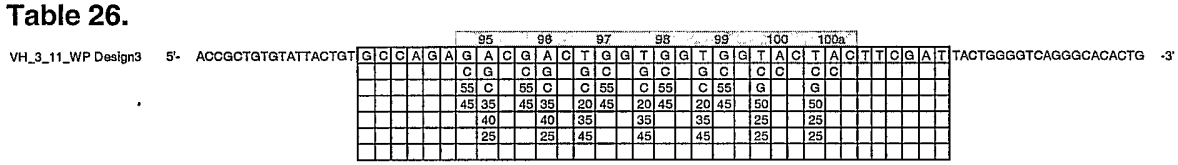


Table 27. Design 4 length 11

VH_CDR3-11	93	94	95	96	97	98	99	100	100a	100b	101
	A	R	X-G/H	X-G/H	X-Y	X-Y	X-Y/H/D/N	X-Y	X-Y	F	D

For design 4, serine is used as the walkthrough/extended walkthrough amino acid.

For positions 95 and 96, glycine (GGC) is the base amino acid, with histidine (CAC) as the required co-product, serine (AGC) is the walkthrough codon. Therefore, A, G and C are used in the first position, G in the second position, and C in the third position.

For position 97, 98, 100 and 100a, tyrosine TAC is the base amino acid, and TCC is the walkthrough codon, yielding T in the first position, A and C in the second position, and C in the final position.

For position 99, tyrosine (TAC) is the base amino acid, and histidine (CAC), aspartate (GAC), and asparagine (AAC) are the required co-products. TCC is utilized for the serine codon. Therefore, A, C, G and T are used in the first position, A and C are used in the second position, and C is used in the third position.

Doping is performed to favor the base amino acid, and the flanking regions are added to yield the following oligonucleotide mixture:

Table 28.

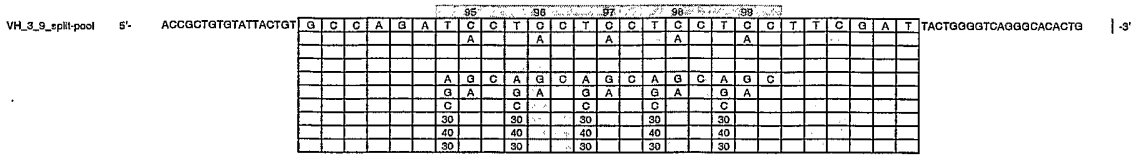
VH_3_11_WS Design 4	5'-	ACCGCTGTGTATTACTGT	95	96	97	98	99	100	100a	TACTGGGGTCAGGGCACACTG	-3'
			G	G	T	T	C	T	T		
				A	A	A		A	A		
			C 55	C 55	40 55	40 55	G	40 55	40 55		
			45 45	45 45	60 45	60 45	A	60 45	60 45		
			30	30			40				
			25	25			20				
							20				
							20				

EXAMPLE 11
SPLIT POOL DESIGN MUTAGENSIS/OLIGONUCLEOTIDE SYNTHESIS

For identifying interesting antigen binding region/antigen contacts, tyrosines and glycines can be incorporated at all residue positions desired. In order to introduce these residues at every position without producing unwanted co-products such as the amber stop codon, an alternate oligonucleotide synthesis procedure can be utilized where pools of codons are synthesized separately then combined and split for the following round of synthesis (E A Peters, P J Schatz, S S Johnson, and W J Dower , J Bacteriol. 1994 July; 176(14): 4296–4305.). In this process, two pools are utilized: the first pool utilizes the codon TMC, encoding Y and S, and the second pool utilizes the codon VRC, encoding

H,S,R,N, G and D. These pools, therefore allow a hydrophobic contribution by tyrosine, and multiple polar contributions and glycine with the second pool. In this split pool design, all diversity positions that are noted with an X in the CDRH3 diversity tables (Figure 12) can contain split pools of these codons. This example shows the codon sets utilized for VH3 CDR length 9 as shown below:

Table 29.



The first pool encodes Y and S at a 50-50 ratio. However, the second pool is doped to increase glycine incorporation to 15% after pooling. The tyrosine pool is encoded at 1/3 the size of the histidine pool to obtain a more balanced ratio of amino acids.

In order to produce the defined mixture of amino acids, four oligonucleotide columns are utilized. First, on all four columns, the fixed 3′ portion of the oligonucleotides are synthesized as defined by the flanking regions and the fixed portion of the CDRH3 shown above. For position 99 in the example sequence above, the first column synthesizes the codon TMC (CMT in the 3′-5′ DNA synthesis). The remaining three columns synthesize the codon VRC (CRV in the 3′-5′ DNA synthesis) utilizing the nucleotide ratios outlined above. After the three nucleotides are coupled, all four columns are opened, the synthesis support is removed by washing with acetonitrile, and the resins are pooled. After mixing, the resin is placed in equal portions to the four columns. At this point, the next position, position 98, is synthesized. One column synthesizes the codon TMC as described above, and three columns synthesize the VRC mixture. The resin is pooled, mixed and reapportioned as described for position 99. This process is repeated for position 97, 96, and 95. At this point, the 5′ fixed and flanking region is added to all four columns, and the resulting oligonucleotide mixture from all four columns can be pooled together and incorporated into an antibody template utilizing a mutagenesis process such as Kunkel mutagenesis.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

Claims

1. A library of polynucleotides encoding antibody binding regions comprising,
a) one or more framework regions selected according to the following criteria:
i) each framework region comprises a germline sequence at or above a defined threshold frequency in expressed antibodies against a predetermined antigen class and, optionally
b) one or more CDR regions.

2. The library of claim 1, wherein antibody binding region is selected from the group consisting of an antibody, an antibody light chain (VL), an antibody heavy chain (VH), and a single chain antibody (scFv).

3. The library of claim 1, wherein the library comprises one or more CDR regions.

4. The library of claim 3, wherein the one or more CDR regions is selected by the following criteria:
i) the CDR region comprises a length determined by the canonical class of the selected framework region(s) of a), and
ii) the CDR region comprises an amino acid residue at each position or a subset of positions within the CDR that is selected from a set of amino acids occurring at or above a threshold frequency found at the corresponding position in the corresponding naturally-occurring CDR region(s),
wherein the framework regions of a) and CDR regions of b) together form an antibody binding region.

5. The library of claim 3, wherein the one or more CDR regions is selected by the following criteria:
i) the CDR region comprises a length determined by the canonical class of the selected framework region(s) of a), and
ii) the CDR region comprises an amino acid residue at each position within the CDR which is represented by an amino acid residue selected from the group consisting of,
a conserved amino acid that occurs in both the germline CDR and expressed CDR region;
a semi-conserved amino acid that is conserved in the germline but is variable in the expressed CDR region;

and a non-conserved amino acid that is variable in the germline and variable in the expressed CDR region;

wherein the framework regions of a) and CDR regions of b) together form an antibody binding region.

6. The library of claim 1, wherein the framework region is selected from the group consisting of a light chain framework region and a heavy chain framework region.

7. The library of claim 6, wherein the light chain framework regions are derived from a light chain human framework clone selected from the group consisting of V_κI-L1, V_κI-L5, V_κIII-A27, V_κIII-L6, V_κIII-L20, V_λ1-1b, V_λ1-1c, V_λ2-2a2, and V_λ3-3l, V_λ3-3r.

8. The library of claim 6, wherein the heavy chain framework regions are derived from a heavy chain human framework clone selected from the group consisting of 1-e, 3-07, 3-11, 3-21, 3-23, 3-30.5, 3-33, 3-48, and 3-74.

9. The library of claim 6, wherein the heavy chain framework comprises a J segment sequence selected from the group consisting of J_H1, J_H2, J_H3, J_H4, J_H5, and J_H6.

10. The library of claim 1, wherein the framework region has a threshold frequency of occurrence of about 10% to about 100%.

11. The library of claim 1, wherein the predetermined antigen class is a class of antigens selected from the group consisting of proteins, peptides, small molecules, polysaccharides, and polynucleotides.

12. The library of claim 1, wherein the CDR regions are selected from the group consisting of CDR-H1, CDR-H2, CDR-H3, CDR-L1, CDR-L2, and CDR-L3.

13. The library of claim 12, wherein the CDR-H1 length determined by the CDR definition chosen relative to the given canonical structure is selected from the group consisting of length five, six, seven and eight.

14. The library of claim 13, wherein the CDR-H1 region is of length five according to the CDR definition based on Kabat.

15. The library of claim 14, wherein CDR-H1 comprises the amino acid sequence SYX_1MX_2 wherein the sequence for X_1 is *A,Y,G,D,S,I,P,R,H, N or G*, and X_2 is *A,Y,G,D,S,I,P,R,H, N,W or G.*

16. The library of claim 13, wherein the CDR-H1 region is of length six according to the CDR definition based on contact determinations.

17. The library of claim 16, wherein VH1 CDR-H1 comprises the amino acid sequence $T/SSYX_1I/MX_2$) wherein X_1 is *A,Y,G,D,S,I,P,R,H, N* or *G*, and X_2 is *A,D,S,I,P,R,Y,H, or N.*

18. The library of claim 16, wherein VH3 CDR-H1 comprises the amino acid sequence $SX_1Y\ X_2MX_3$ wherein X_1 is *A,D,S,I,P,R,Y,H,N or T* and X_2 is *A,D,S,I,P,R,Y,H,N,W or G* and X_3 is *A,D,S,I,P,R,Y,H,N*

19. The library of claim 12, wherein the CDR-H2 length is determined by the CDR definition chosen relative to the given canonical structure selected from the group consisting of length seventeen, twelve, and thirteen.

20. The library of claim 19, wherein the CDR-H2 region is of length seventeen according to CDR definition based on Kabat.

21. The library of claim 20, wherein CDR-H2 comprises the amino acid sequence $X_1I\ X_2X_3X_4GGX_4X_6YYADSVKG$ wherein X_1 is *A,D,S,I,P,R,Y,H,N, G or W*, X_2 is *A,D,S,I,P,R,Y,H or N* , X_3 is *A,D,S,I,P,R,Y,H or N*, X_4 is *A,D,S,I,P,R,Y,H,N or G*, X_5 is *A,D,S,I,P,R,Y,H or N* and X_6 *is A,D,S,I,P,R,Y,H,N or T.*

22. The library of claim 19, wherein the CDR-H2 region is of length thirteen according to the CDR definition based on contact determinations.

23. The library of claim 22, wherein VH1 CDR-H2 comprises the amino acid sequence of $WMGX_1I\ X_2PX_3X_4G\ X_5T/AN$ wherein X_1 is A,D,S,I,P,R,Y,H,N, G or W, X_2 is A,D,S,I,P,R,Y,H or N , X_3 is A,D,S,I,P,R,Y,H,N,G or M, X_4 is A,D,S,I,P,R,Y,H,N,F or G, X_5 is A,D,S,I,P,R,Y,H,N or T.

24. The library of claim 22, wherein VH3 CDR-H2 comprises the amino acid sequence of $WVS/AX_1IS\ X_2X_3GX_4X_5\ X_6Y$ wherein X_1 is A,D,S,I,P,R,Y,H,N,V,G or T, X_2 is A,D,S,I,P,R,Y,H,N,G,Q,W or F , X_3 is A,D,S,I,P,R,Y,H, or N, X_4 is

A,D,S,I,P,R,Y,H,N,G or T, X_5 is A,D,S,I,P,R,Y,H,N,T or K and X_6 is A,D,S,I,P,R,Y,H,N or F .

25. The library of claim 12, wherein the CDR-H3 length is selected from the group consisting of length nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen and eighteen.

26. The library of claim 25, wherein the CDR-H3 region is of length nine.

27. The library of claim 26, wherein CDR-H3 comprises the amino acid sequence of $ARX_1X_2X_3X_4X_5FD$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E,L or V* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L or* Q and X_3 is *A,D,S,I,P,R,Y,H,N,G,V,T,L or Q* and X_4 is *A,D,S,I,P,R,Y,H,N,G,W or L* X_5 is *A,D,S,I,P,R,Y,H,N,G,L,T or V.*

28. The library of claim 25, wherein the CDR-H3 region is of length ten.

29. The library of claim 28, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E,L,V or M* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L,V,K or Q* and X_3 is *A,D,S,I,P,R,Y,H,N,G,T,E,L,V,Q orW* and X_4 is *A,D,S,I,P,R,Y,H,N,G,L,V,W or Q* and X_5 is *A,D,S,I,P,R,Y,H,N,G,T,E,W or L* and X_6 is *A,D,S,I,P,R,Y,H,N,G,L or F.*

30. The library of claim 25, wherein the CDR-H3 region is of length eleven.

31. The library of claim 30, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6 X_7FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,K,G,E or L* and X_2 is *A,D,S,I,P,R,Y,H,N,T,G,F,L or V* and X_3 is *A,D,S,I,P,R,Y,H,N,G,V or T* and X_4 is *A,D,S,I,P,R,Y,H,N,G,T or W* and X_5 is *A,D,S,I,P,R,Y,H,N,G,T or L* and X_6 is *A,D,S,I,P,R,Y,H,N,G or W* and X_7 is *A,D,S,I,P,R,Y,H,N,G,W,F or L).*

32. The library of claim 25, wherein the CDR-H3 region is of length twelve.

33. The library of claim 32, wherein CDR-H3 comprises the amino acid sequence of $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E or V* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L,Q or T* and X_3 is *A,D,S,I,P,R,Y,H,N,G,L,T,V or W* and X_4 is *A,D,S,I,P,R,Y,H,N,G,W,L or V* and X_5 is *A,D,S,I,P,R,Y,H,N,G,V,F,T or L* and X_6 is *A,D,S,I,P,R,Y,H,N,G,L,T or E* and X_7 is *A,D,S,I,P,R,Y,H,N,G,T,W or F* and X_8 is *A,D,S,I,P,R,Y,H,N,G,F,T or W.*

34. The library of claim 25, wherein the CDR-H3 region is of length thirteen.

35. The library of claim 34, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E,V,L or K* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L,Q or K* and X_3 is *A,D,S,I,P,R,Y,H,N,G,L,V,K or M* and X_4 is *A,D,S,I,P,R,Y,H,N,G,T,L or V* and X_5 is *A,D,S,I,P,R,Y,H,N,G,T,W,L or Q* and X_6 is *A,D,S,I,P,R,Y,H,N,G,V,L,E or T* and X_7 is *A,D,S,I,P,R,Y,H,N,L,V,T,W or G* and X_8 is *A,D,S,I,P,R,Y,H,N,G or F* and X_9 is *A,D,S,I,P,R,Y,H,N,G,F,W or T.*

36. The library of claim 25, wherein the CDR-H3 region is of length fourteen.

37. The library of claim 36, wherein CDR-H3 comprises the amino acid sequence of $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9 X_{10}FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E, or V* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L,T,Q or K* and X_3 is *A,D,S,I,P,R,Y,H,N,G,L,V,T or E* and X_4 is *A,D,S,I,P,R,Y,H,N,G,L,F,T or Q* and X_5 is *A,D,S,I,P,R,Y,H,N,G,V,T or L* and X_6 is *A,D,S,I,P,R,Y,H,N,G,T,L or V* and X_7 is *A,D,S,I,P,R,Y,H,N,G,T,L,E or V* and X_8 is *A,D,S,I,P,R,Y,H,N,G,T,F,E or L* and X_9 is *A,D,S,I,P,R,Y,H,N,G,W or T* and X_{10} is *A,D,S,I,P,R,Y,H,N,G,F,W or L).*

38. The library of claim 25, wherein the CDR-H3 region is of length fifteen.

39. The library of claim 38, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9 X_{10} X_{11}FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E,V or T* and X_2 is *A,D,S,I,P,R,Y,H,N,G,W,L,T or V* and X_3 is *A,D,S,I,P,R,Y,H,N,G,E,T,F,L or W* and X_4 is *A,D,S,I,P,R,Y,H,N,G,E,C,T or F* and X_5 is *A,D,S,I,P,R,Y,H,N,G,T,W or L* and X_6 is *A,D,S,I,P,R,Y,H,N,G,T or E* and X_7 is *A,D,S,I,P,R,Y,H,N,G,T,V or W* and X_8 is *A,D,S,I,P,R,Y,H,N,G,T,L,F,V,M or W* and X_9 is *A,D,S,I,P,R,Y,H,N,G,V,C or K* and X_{10} is *A,D,S,I,P,R,Y,H,N,G,W or Q* and X_{11} is *A,D,S,I,P,R,Y,H,N,G,W,F or L.*

40. The library of claim 25, wherein the CDR-H3 region is of length sixteen.

41. The library of claim 40, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12}FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,L,V or E* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L,V or E* and X_3 is *A,D,S,I,P,R,Y,H,N,G,T or L* and X_4 is *A,D,S,I,P,R,Y,H,N,G,L,T or E* and X_5 is *A,D,S,I,P,R,Y,H,N,G,V,F,T or M* and X_6 is *A,D,S,I,P,R,Y,H,N,G,T,F or W* and X_7 is *A,D,S,I,P,R,Y,H,N,G,T,VL or E* and X_8 is *A,D,S,I,P,R,Y,H,N,G,T,E or W* and X_9 is *A,D,S,I,P,R,Y,H,N,G,L,F or W* and X_{10} is

A,D,S,I,P,R,Y,H,N,G,L,T or F and X_{11} is *A,D,S,I,P,R,Y,H,N,G,W or T* and X_{12} is *A,D,S,I,P,R,Y,H,N,G,W or F.*

42. The library of claim 25, wherein the CDR-H3 region is of length seventeen.

43. The library of claim 42, wherein CDR-H3 comprises the amino acid sequence $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13}FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,V,E or L* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L or Q* and X_3 is *A,D,S,I,P,R,Y,H,N,G,L,T,V or M* and X_4 is *A,D,S,I,P,R,Y,H,N,G,V or C* and X_5 is *A,D,S,I,P,R,Y,H,N,G,V or T* and X_6 is *A,D,S,I,P,R,Y,H,N,G,F,V or T* and X_7 is *A,D,S,I,P,R,Y,H,N,G,W,T,V or F* and X_8 is *A,D,S,I,P,R,Y,H,N,G,L or V* and X_9 is *A,D,S,I,P,R,Y,H,N,G,V,F,L or C* and X_{10} is *A,D,S,I,P,R,Y,H,N,G,F or L* and X_{11} is *A,D,S,I,P,R,Y,H,N,G,L,V or C* and X_{12} is *A,D,S,I,P,R,Y,H,N or G* and X_{13} is *A,D,S,I,P,R,Y,H,N,G or W.*

44. The library of claim 25, wherein the CDR-H3 region is of length eighteen.

45. The library of claim 44, wherein CDR-H3 comprises the amino acid sequence of $ARX_1X_2X_3X_4X_5$ X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14}FD wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,E,V or L* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L or K* and X_3 is *A,D,S,I,P,R,Y,H,N,G,T,V,L or F* and X_4 is *A,D,S,I,P,R,Y,H,N,G,V or T* and X_5 is *A,D,S,I,P,R,Y,H,N,G,C,F,K,L or M* and X_6 is *A,D,S,I,P,R,Y,H,N,G,V,F,T,W or C* and X_7 is *A,D,S,I,P,R,Y,H,N,G,T,W or F* and X_8 is *A,D,S,I,P,R,Y,H,N,G,W,V or F* and X_9 is *A,D,S,I,P,R,Y,H,N,G,L or V* and X_{10} is *A,D,S,I,P,R,Y,H,N,G,C,L or F* and X_{11} is *A,D,S,I,P,R,Y,H,N,F,G or W* and X_{12} is *A,D,S,I,P,R,Y,H,N,G,T,L or F* and X_{13} is *A,D,S,I,P,R,Y,H,N,G or W* and X_{14} is *A,D,S,I,P,R,Y,H,N,G,W or T.*

46. The library of claim 12, wherein the VL (kappa) CDR-L1 length is determined by the CDR definition chosen relative to the given canonical structure selected from the group consisting of length seven and eight.

47. The library of claim 46, wherein the VK-I CDR-L1 region is of length seven according to the contact CDR definition.

48. The library of claim 47, wherein CDR-L1 comprises the amino acid sequence SX_1X_2LA/NWY wherein X_1 is *A,D,S,I,P,R,Y,H,N,T or K* and X_2 is *A,D,S,I,P,R,Y,H,N or W.*

49. The library of claim 46, wherein the VK-III CDR-L1 region is of length seven according to the contact CDR definition.

50. The library of claim 49, wherein CDR-L1 comprises the amino acid sequence of SSN/YLAWY.

51. The library of claim 46, wherein the VK-III CDR-L1 region is of length eight according to the contact CDR definition.

52. The library of claim 51, wherein CDR-L1 comprises the amino acid sequence SS/NX1YLAWY wherein X_1 is *A,D,S,I,P,R,Y,H,N, or T.*

53. The library of claim 12, wherein the VL (kappa) CDR-L2 length is determined by the CDR definition chosen relative to the given canonical structure selected from the group consisting of length ten.

54. The library of claim 53, wherein the VK-I CDR-L2 region is of length ten according to the contact CDR definition.

55. The library of claim 54, wherein CDR-L2 comprises the amino acid sequence of $LLIYX_1ASX_2LQ/E$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,K or G* and X_2 is *A,D,S,I,P,R,Y,H,N or T.*

56. The library of claim 53, wherein the VK-III CDR-L2 region is of length ten according to the contact CDR definition.

57. The library of claim 56, wherein CDR-L2 comprises the amino acid sequence $LLIYG/DASX_1RA$ wherein X_1 is *A,D,S,I,P,R,Y,H,N or T*).

58. The library of claim 12, wherein the VL (kappa) CDR-L3 length is determined by the CDR definition selected from the group consisting of length eight and nine.

59. The library of claim 58, wherein the CDR-L3 region is of length eight according to the contact CDR definition.

60. The library of claim 59, wherein CDR-L3 comprises the amino acid sequence $QQYX_1X_2X_3PX_4$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,G,T or L* and X_2 is *A,D,S,I,P,R,Y,H,N,T,Q or G* and X_3 is *A,D,S,I,P,R,Y,H,N,T,L,W or F* and X_4 is *A,D,S,I,P,R,Y,H,N,L,W or F.*

61. The library of claim 58, wherein the CDR-L3 region is of length nine according to the contact CDR definition.

62. The library of claim 61, wherein CDR-L3 comprises the amino acid sequence $QQYX_1X_2X_3PPX_4$ wherein X_1 is *A,D,S,I,P,R,Y,H,N or G* and X_2 is *A,D,S,I,P,R,Y,H,N,T or G* and X_3 is *A,D,S,I,P,R,Y,H,N,W or T* and X_4 is *A,D,S,I,P,R,Y,H,N,W,L or T.*

63. The library of claim 12, wherein the Vl (lambda) CDR-L1 length is determined by the CDR definition chosen relative to the given canonical structure selected from the group consisting of length seven, nine and ten.

64. The library of claim 63, wherein the Vλ-1 CDR-L1 region is of length nine according to the contact CDR definition.

65. The library of claim 64, wherein CDR-L1 comprises the amino acid sequence $IGX_1NX_2V\ X_3WY$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,T or G* and X_2 is *A,D,S,I,P,R,Y,H,N,T or F* and X_3 is *A,D,S,I,P,R,Y,H or N.*

66. The library of claim 63, wherein the Vλ-2 CDR-L1 region is of length ten according to the contact CDR definition.

67. The library of claim 66, wherein CDR-L1 comprises the amino acid sequence of $VGX_1YNYVSWY$ wherein X_1 is *A,D,S,I,P,R,Y,H,N or G*)

68. The library of claim 63, wherein the Vλ-3 CDR-L1 region is of length seven according to the contact CDR definition.

69. The library of claim 68, wherein CDR-L1 comprises the amino acid sequence of $X_1X_2X_3A/V\ X_4WY$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,K or T* and X_2 is *A,D,S,I,P,R,Y,H,N,K,Q or E* and X_3 is *A,D,S,I,P,R,Y,H,N or F* and X_4 is *A,D,S,I,P,R,Y,H,N or C.*

70. The library of claim 12, wherein the VL (lambda) CDR-L2 length is determined by the CDR definition chosen relative to the given canonical structure selected from the group consisting of length ten.

71. The library of claim 70, wherein the Vλ-1 CDR-L2 region is of length ten according to the contact CDR definition.

72. The library of claim 71, wherein CDR-L1 comprises the amino acid sequence LLIYX$_1$NN/SX$_2$RP wherein X$_1$ is *A,D,S,I,P,R,Y,H,N,G or E* and X$_2$ is *A,D,S,I,P,R,Y,H,N,Q or K.*

73. The library of claim 70, wherein the Vλ-2 CDR-L2 region is of length ten according to the contact CDR definition.

74. The library of claim 73, wherein CDR-L1 comprises the amino acid sequence LM/IIYE/DVX$_1$X$_2$RP wherein X$_1$ is *A,D,S,I,P,R,Y,H,N or T* and X$_2$ is *A,D,S,I,P,R,Y,H,N or K.*

75. The library of claim 70, wherein the Vλ-3 CDR-L2 region is of length ten according to the contact CDR definition.

76. The library of claim 75, wherein CDR-L1 comprises the amino acid sequence LVI/VYX$_1$DX$_2$X$_3$RP wherein X$_1$ is *A,D,S,I,P,R,Y,H,N,Q,E,K or G* and X$_2$ is *A,D,S,I,P,R,Y,H,N or T* and X$_3$ is *A,D,S,I,P,R,Y,H,N,K or E.*

77. The library of claim 12, wherein the VL (lambda) CDR-L3 length is determined by the CDR definition selected from the group consisting of length eight, nine, ten and eleven.

78. The library of claim 77, wherein the CDR-L3 region is of length eight according to the contact CDR definition.

79. The library of claim 78, wherein CDR-L3 comprises the amino acid sequence QS/AWDX$_1$SX$_2$X$_3$ wherein X$_1$ is *A,D,S,I,P,R,Y,H,N or G* and X$_2$ is *A,D,S,I,P,R,Y,H,N,T,L or G* and X$_3$ is *A,D,S,I,P,R,Y,H,N,V,W,Q or L.*

80. The library of claim 77, wherein the CDR-L3 region is of length nine according to the contact CDR definition.

81. The library of claim 80, wherein CDR-L3 comprises the amino acid sequence QS/AYD/A$X_1SX_2X_3X_4$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,G or T* and X_2 is *A,D,S,I,P,R,Y,H,N,G,L or T* and X_3 is *A,D,S,I,P,R,Y,H,N,T or L* and X_4 is *A,D,S,I,P,R,Y,H,N,V,W,L,F or G.*

82. The library of claim 77, wherein the CDR-L3 region is of length ten according to the contact CDR definition.

83. The library of claim 82, wherein CDR-L3 comprises the amino acid sequence QS/AWDX_1SL/S$X_2X_3X_4$ wherein X_1 is *A,D,S,I,P,R,Y,H,N,T or G* and X_2 is *A,D,S,I,P,R,Y,H,N or T* and X_3 is *A,D,S,I,P,R,Y,H,N,G,L or V* and X_4 is *A,D,S,I,P,R,Y,H,N,V,W or G.*

84. The library of claim 77, wherein the CDR-L3 region is of length eleven according to the contact CDR definition.

85. The library of claim 84, wherein CDR-L3 comprises the amino acid sequence QS/AWDX_1SL/S$X_2X_3X_4X_5$ wherein X_1 is *A,D,S,I,P,R,Y,H,N or G* and X_2 is *A,D,S,I,P,R,Y,H,N or T* and X_3 is *A,D,S,I,P,R,Y,H,N, or L* and X_4 is *A,D,S,I,P,R,Y,H,N,V,L or F* and X_5 is *A,D,S,I,P,R,Y,H,N,V,W or G.*

86. The library of claim 1, wherein the CDR region comprises amino acid residues naturally occurring at a threshold frequency of about 10% to about 100%.

87. The library of claim 1, wherein the CDR regions have a diversity wherein one residue within the CDR is altered using look-though mutagenesis.

88. The library of claim 1, wherein the CDR regions have a diversity wherein more than one residue within the CDR is altered using walk-through mutagenesis.

89. The library of claim 1, wherein the antibody binding regions further comprise one or more amino acid substitutions corresponding to a naturally occurring somatic mutation.

90. The library of claim 1, wherein the library is an expression library.

91. The library of claim 90, wherein the expression library is selected from the group consisting of a ribosome display library, a polysome display library, a phage display library, a bacterial expression library, and a yeast display library.

92. The library of claim 1, wherein the library comprises a diversity of antibody binding regions selected from the group consisting of at least about 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, $10^{12}$and 10^{13}.

93. The library of claim 1, produced by synthesizing polynucleotides encoding one or more framework regions and one or more CDR regions wherein the polynucleotides are predetermined, wherein the polynucleotides encoding said regions further comprise sufficient overlapping sequence whereby the polynucleotide sequences, under polymerase chain reaction (PCR) conditions, are capable of assembly into polynucleotides encoding complete antibody binding regions.

94. The library of claim 93, wherein the polynucleotides encoding the defined CDR regions are mutagenized using a mutagenesis selected from the group consisting of walk-through-mutagenesis (WTM), extended walk-through-mutagenesis, look-through-mutagenesis (LTM), and a combination thereof.

95. A method of producing the library of claim 1 comprising,
synthesizing polynucleotides encoding one or more framework regions selected according to the following criteria:

i) each framework region occurs at or above a threshold frequency in expressed antibodies against a predetermined antigen class and, optionally

ii) each framework region occurs at or above a threshold frequency in a germline antibody sequence; and synthesizing polynucleotides encoding one or more defined CDR regions having a predetermined diversity, wherein the polynucleotides encoding said regions further comprise sufficient overlapping polynucleotide sequence whereby the polynucleotide sequences, under polymerase chain reaction (PCR) conditions, are capable of assembly into polynucleotides encoding complete antibody binding regions.

96. The method of claim 95, wherein the predetermined diversity of the defined CDRs is designed according to the following criteria:

i) the length of the CDR regions is determined by the resultant canonical structure of the selected framework regions, and

ii) the CDR regions comprise a diversity of amino acids residues at a threshold frequency of occurrence found naturally occurring within the defined CDR region.

97. The method of claim 96, wherein the predetermined diversity is introduced using mutagenesis selected from the group consisting of walk-through-mutagenesis (WTM), extended walk-through-mutagenesis, look-through-mutagenesis (LTM), and a combination thereof.

98. A method of identifying a polypeptide having a desired binding affinity comprising,

expressing the expression library of claim 87 to produce antibody binding regions, and

screening the antibody binding regions to select for an antibody binding region having a desired binding affinity.

99. The method of claim 98, wherein the screening comprises,

contacting the antigen binding region with a target substrate, the antibody binding region being associated with the polynucleotide encoding the antibody binding region.

100. The method of claim 98, wherein the method further comprises the step of identifying the polynucleotide that encodes the selected antibody binding region.

101. The method of claim 29, wherein the polynucleotide is associated with the antibody binding region using an expression display selected from the group consisting of phage display, bacteria display, and yeast display.

102. An antibody binding region identified according to the method of claim 99 or 100.

103. The method of claim 95 or 98, wherein one or more steps is computer-assisted.

104. A medium suitable for use in an electronic device having instructions for carrying out one or more steps of the method of claim 103.

105. A device for carrying out one or more steps of the method of claim 104.

106. A library of polynucleotides encoding antibody binding regions comprising,

a) one or more framework regions having a sequence set forth in any of the figures or tables; and

b) one or more CDR regions having a sequence set forth in any of the figures or tables

wherein the framework regions of a) and CDR regions of b) together form an antibody binding region against a predetermined antigen class.

Fig. 1

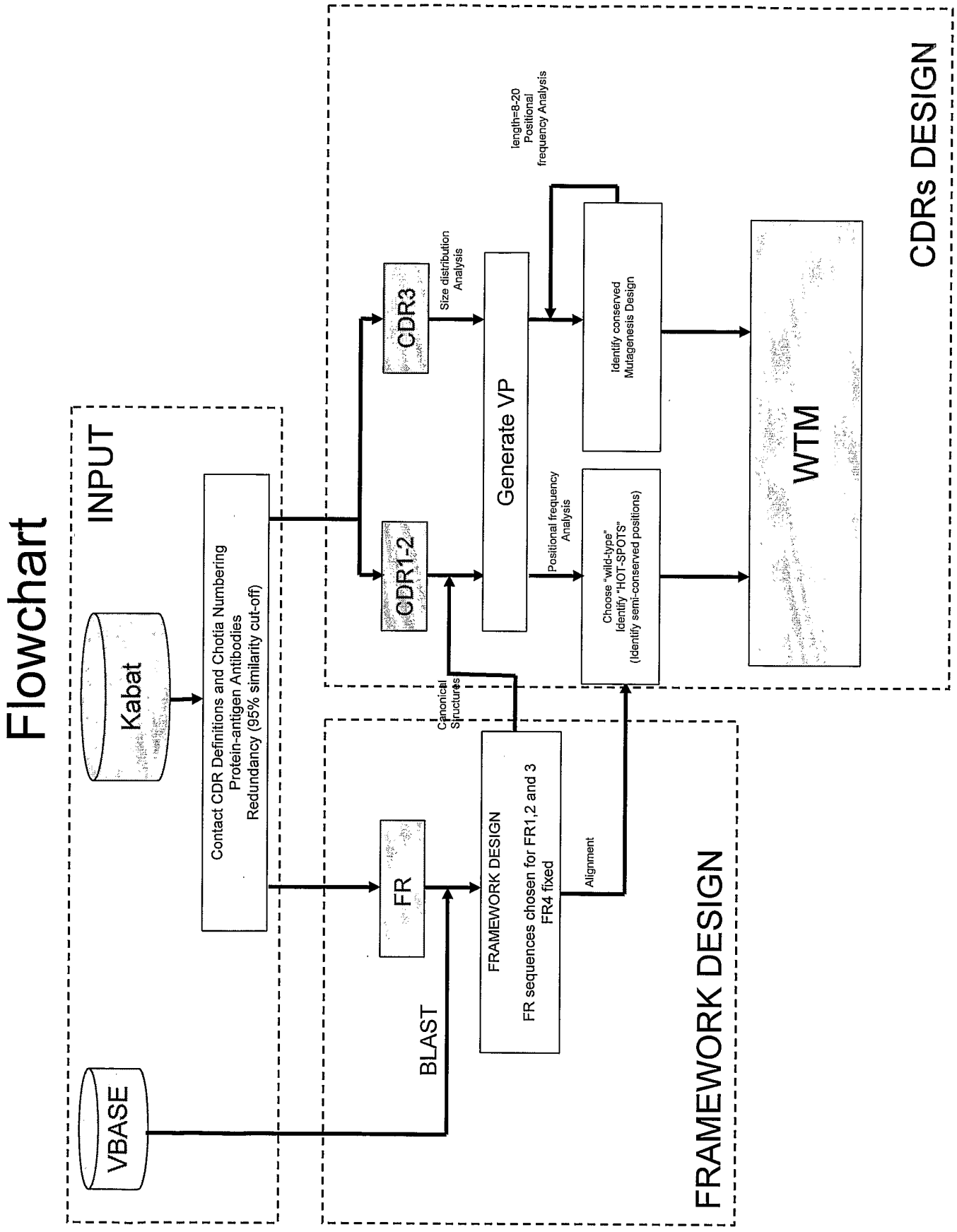

Flowchart
INPUT
Kabat
VBASE
Contact CDR Definitions and Chotia Numbering
Protein-antigen Antibodies
Redundancy (95% similarity cut-off)
BLAST
FR
FRAMEWORK DESIGN
FR sequences chosen for FR1,2 and 3
FR4 fixed
Alignment
Canonical Structures
FRAMEWORK DESIGN
CDR1-2
CDR3
Size distribution Analysis
length=8-20
Positional frequency Analysis
Generate VP
Positional frequency Analysis
Identify conserved Mutagenesis Design
Choose "wild-type"
Identify "HOT-SPOTS"
(Identify semi-conserved positions)
WTM
CDRs DESIGN

Fig. 2

IDENTIFY & SELECT FROM VBASE

Database of germline families
51 functional VH segments
7 families: VH1-7
40 functional Vκ segments
7 families: Vk I-VII
31 functional Vλ segments
10 families: Vλ 1-10

IDENTIFY & SELECT FROM KABAT & KABATMAN DATABASES

Kabat	Kabatman
5977 VH sequence	3319 VH
2374 Vk sequences	1330 Vk
2012 Vλ sequences	1265 Vλ

INPUT

VBASE database
Full database
Kabatman database
FILTERS:
Kabat Loop Definitions and Numbering
Human sequence with protein/peptide-antigen annotation
Redundancy filter (90% homology tolerance)
Filtered Dataset:
600 VH
319 Vk
156 Vλ

FRAMEWORK SELECTION STRATEGY

Blast Analysis of germline frameworks (VBASE) in rearranged genes (filtered-kabatman)
Select most frequent framework families
Use most common framework 4
Identify "Hot-spots" for somatic hypermutations for future affinity maturation (WTM)
Blast Analysis of germline frameworks (VBASE) in rearranged genes (filtered-kabatman)
Select most frequent framework families
Use most common framework 4
Identify "Hot-spots" for somatic hypermutations for future affinity maturation (WTM)

CDR Design Strategy

Length of CDR1 & 2 dictated by canonical structures of selected frameworks.
Size distribution of CDR3 from frequency analysis of anti-protein/peptide antibody sequences.
Identify conserved positions by frequency analysis of germline (VBASE) and rearranged genes (Kabatman).
Choose highest frequency amino acids as wildtype sequence in non-conserved positions and conduct WTM.
Identify positions to conduct affinity maturation.

Fig. 3

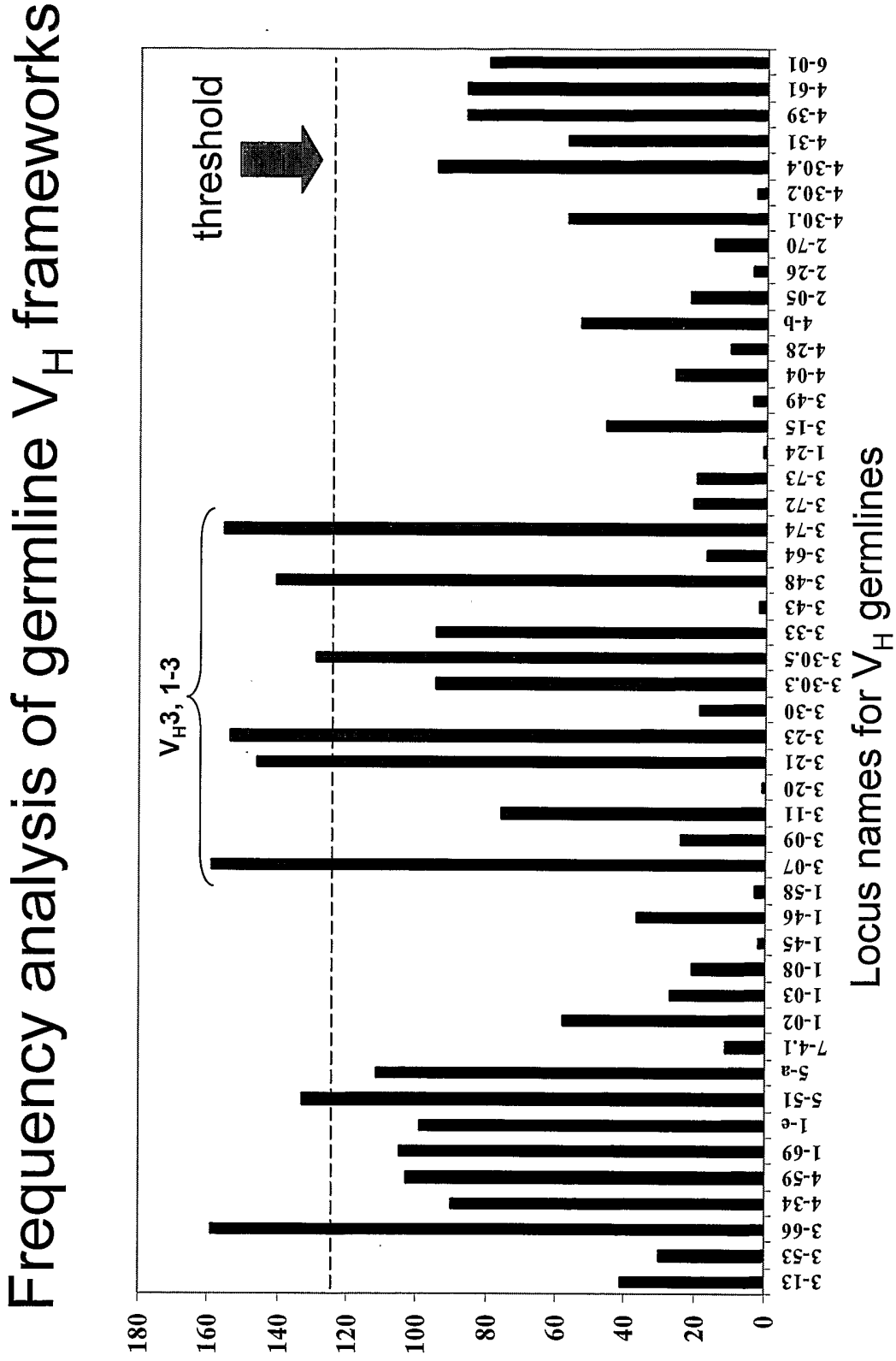

Frequency analysis of germline V_H frameworks
threshold
V_H3, 1-3
Locus names for V_H germlines
180
160
140
120
100
80
60
40
20
0
3-13
3-53
3-66
4-34
4-59
1-69
1-e
5-51
5-a
7-4.1
1-02
1-03
1-08
1-45
1-46
1-58
3-07
3-09
3-11
3-20
3-21
3-23
3-30
3-30.3
3-30.5
3-33
3-43
3-48
3-64
3-74
3-72
3-73
1-24
3-15
3-49
4-04
4-28
4-b
2-05
2-26
2-70
4-30.1
4-30.2
4-30.4
4-31
4-39
4-61
6-01

Fig. 4

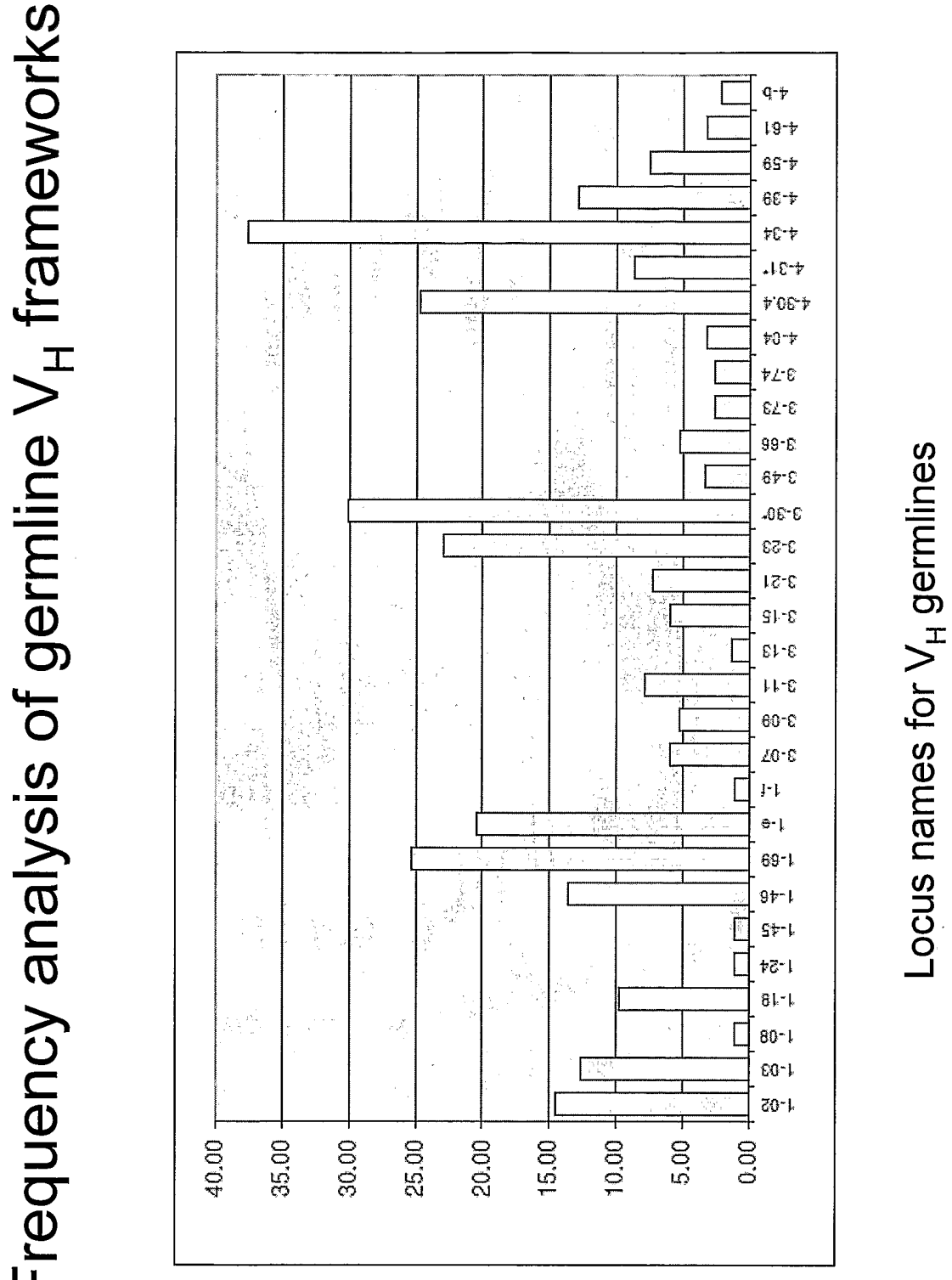

Frequency analysis of germline V_H frameworks
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
1-02
1-03
1-08
1-18
1-24
1-45
1-46
1-69
1-e
1-f
3-07
3-09
3-11
3-13
3-15
3-21
3-23
3-30'
3-49
3-66
3-73
3-74
4-04
4-30.4
4-31'
4-34
4-39
4-59
4-61
4-b
Locus names for V_H germlines

Fig. 5

High Frequency V_H Frameworks

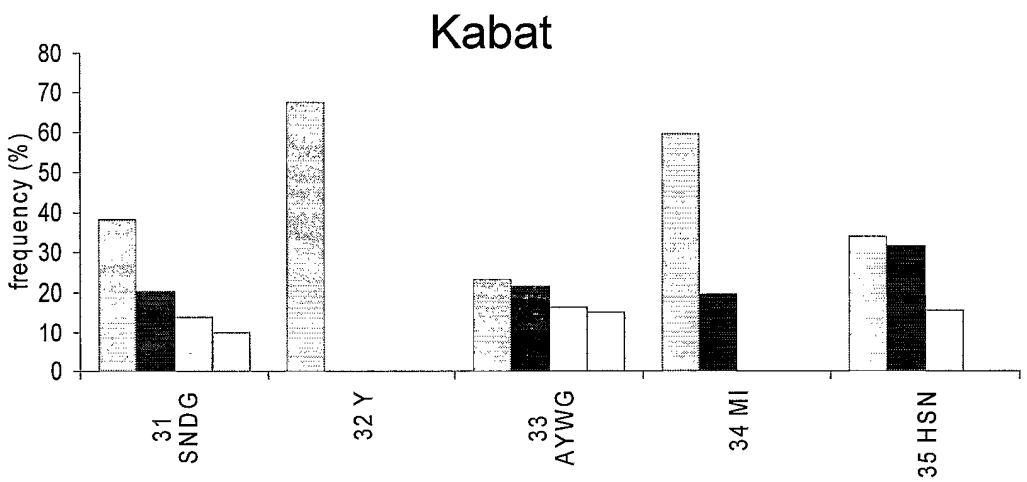
VH

	FR1			CDR1	FR2	CDR2		FR3								CDR3		FR4		
	1		2		3	4	5	6	7	8		9				1 0		1		1
	123456789012345678901234567890				01ab2345	67890123456	789012a345678	9012345678901234567890123456789012abc3456789012								34567890abc1		234567890123		
1-e	QVQLVQSGAEVKKPGSSVKVSCKASGGTF				SS--YAIS	WVRQAPGQGLE	WMGGIIPIFGTAN	YAQKFQGRVTITADESTSTAYMELSSLRSEDTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		
3-07	EVQLVESGGGLVQPGGSLRLSCAASGFTF				SS--YWMS	WVRQAPGKGLE	WVANIKQDGSEKY	YVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		
3-11	QVQLVESGGGLVKPGGSLRLSCAASGFTF				SD--YYMS	WIRQAPGKGLE	WVSYISSSGSTIY	YADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		
3-23	EVQLLESGGGLVQPGGSLRLSCAASGFTF				SS--YAMS	WVRQAPGKGLE	WVSAISGSGGSTY	YADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC								AKDGSGSGYAFD		YWGQGTLVTVSS		
3-33	QVQLVESGGGVVQPGRSLRLSCAASGFTF				SS--YGMH	WVRQAPGKGLE	WVAVIWYDGSNKY	YADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		
4-34	QVQLQQWGAGLLKPSETLSLTCAVYGGSF				SG--YYWS	WIRQPPGKGLE	WIGEI-NHSGSTN	YNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		
4-30.4	QVQLQESGPGLVKPSQTLSLTCTVSGGSI				SSGDYYWS	WIRQPPGKGLE	WIGYI-YYSGSTY	YNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYC								ARDGSGSGYAFD		YWGQGTLVTVSS		

V_H-CDR1

VBASE

```
                ** *
   1-3@3-07  SYWMS
   1-3@3-21  SYSMN
   1-3@3-23  SYAMS
 1-3@3-30.5  SYGMH
   1-3@3-48  SYSMN
   1-3@3-74  SYWMH
```



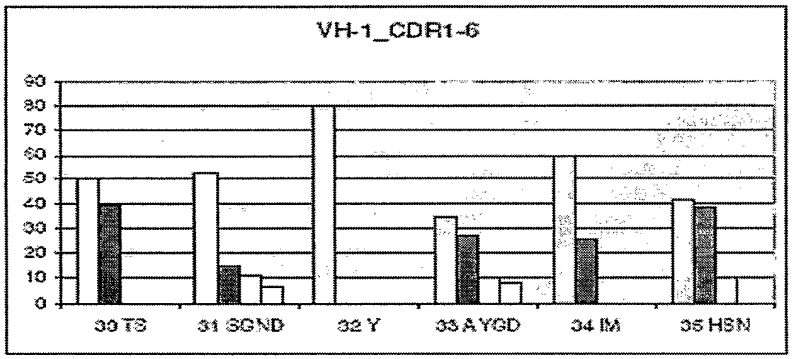
- All CDR1-1 loops have 5 amino acids (Kabat Definition)
- Position 32 is conserved
- Position 31 and 34 are mutated during affinity maturation
- Positions 33 and 35 are mutated

SYXMX

▪INVARIANT POSITIONS
•WTM IN UNIVERSAL LIBRARIES
•WTM IN AFFINITY MATURATION

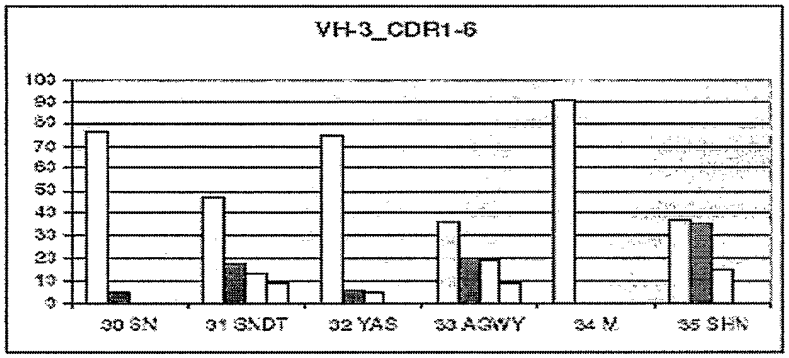
Fig. 6

V_H-CDR1 Sequence Matrices (Contact Definition)



30	31	32	33	34	35
T–S	S	Y	X_A	I–M	X_S

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30			X																X		
31			X																		
32							X														
33	X	X	X	X	X	X	X	X	X							X			X	X	
34				X							X										
35	X	X	X	X	X	X	X	X	X												



30	31	32	33	34	35
S	X_S	Y	X_A	M	X_Z

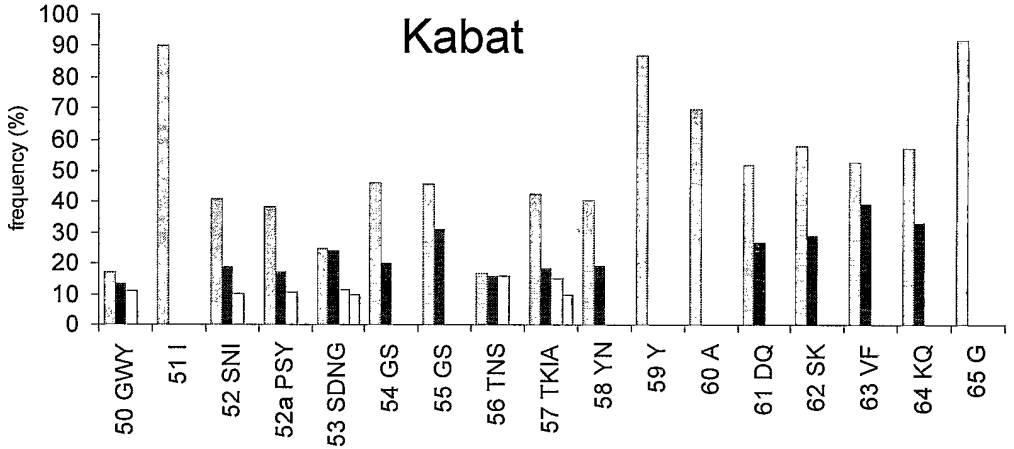
	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30			X																		
31	X	X	X	X	X	X	X	X	X												
32							X														
33	X	X	X	X	X	X	X	X	X												
34											X										
35	X	X	X	X	X	X	X	X	X												

Fig. 7

V_H-CDR2

VBASE

```
                *.  ...    *.*****
1-3@3-07  NIKQDGSEKYYVDSVKG
1-3@3-21  SISSSSSYIYYADSVKG
1-3@3-23  AISGSGGSTYYADSVKG
1-3@3-30.5 VISYDGSNKYYADSVKG
1-3@3-48  YISSSSSTIYYADSVKG
1-3@3-74  RINSDGSSTSYADSVKG
```



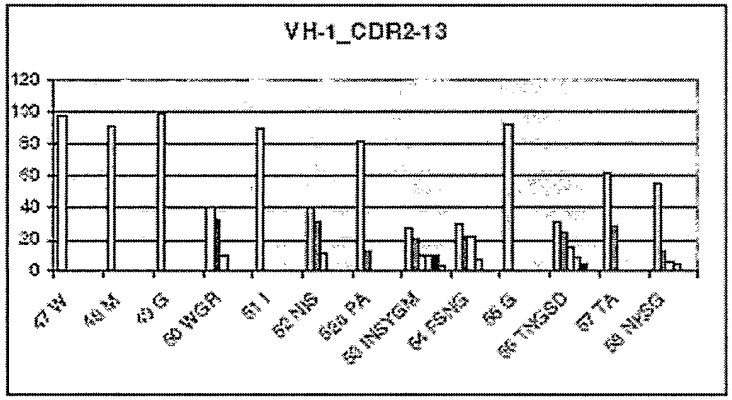
- All CDR2-3,2 loops have 17 amino acids (Kabat Definition)
- Positions 51, 59, 60, and 65 are conserved as single-amino acids
- Positions 54, 55, 58, 61, 62, 63 and 64 are mutated during affinity maturation
- Positions 50, 52, 52a, 53, 56 and 57 are mutated

XIXXXGGXXYYADSVKG

▪INVARIANT POSITIONS
•WTM IN UNIVERSAL LIBRARIES
•WTM IN AFFINITY MATURATION

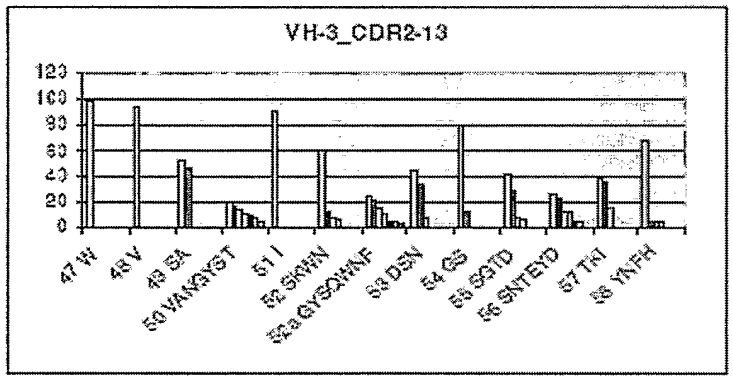
Fig. 8

V_H-CDR2 Sequence Matrices (Contact Definition)



47	48	49	50	51	52	52A	53	54	55	56	57	58
W	M	G	X_C	I	X_H	P	XI	X_3	G	X_E	T-A	N

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
47																					X
48											X										
49																X					
50	X	X	X	X	X	X	X	X	X							X					
51				X																	
52	X	X	X	X	X	X	X	X	X												
52A					X																
53	X	X	X	X	X	X	X	X	X												
54	X	X	X	X	X	X	X	X	X												
55																X					
56	X	X	X	X	X	X	X	X	X										X		
57	X																		X		
58									X												

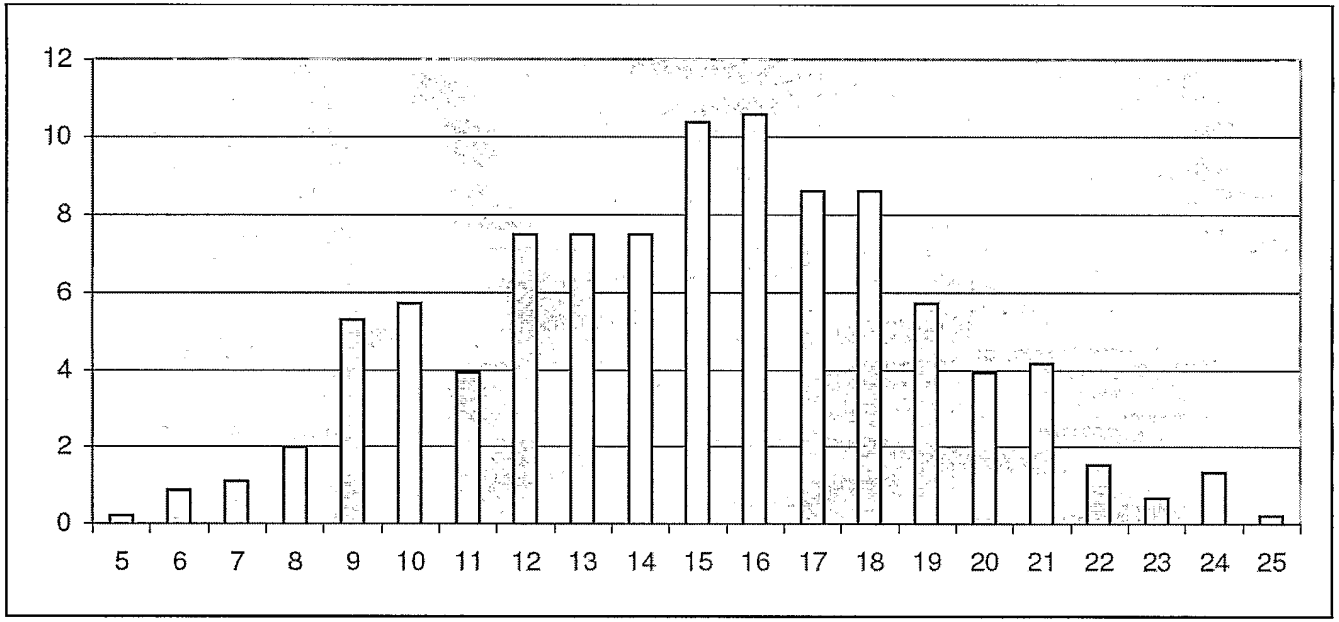


47	48	49	50	51	52	52A	53	54	55	56	57	58
W	V	S-A	X_V	I	S	X_2	X_D	G	X_3	X_E	X_F	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
47																					X
48																				X	
49	X		X																		
50	X	X	X	X	X	X	X	X	X											X	
51				X																	
52			X																		
52A	X	X	X	X	X	X	X	X	X							X					
53	X	X	X	X	X	X	X	X	X											X	
54																X					
55			X													X					
56	X	X	X	X	X	X	X	X	X												
57	X	X	X	X	X	X	X	X	X			X							X		
58							X														

Fig. 9

V_H-CDR3 Length Analysis



- The length variability of V_H CDR3 ranges from 4 to 28 residues (Contact Definition)
- However most (~75%) V_H CDR3s of anti-protein antibodies range from 9 to 18 residues
- All 13 (9, 10,..18) lengths are synthesized separately and pooled before gene assembly in desired ratios

V_H-CDR3 (length 13 Kabat Definition)

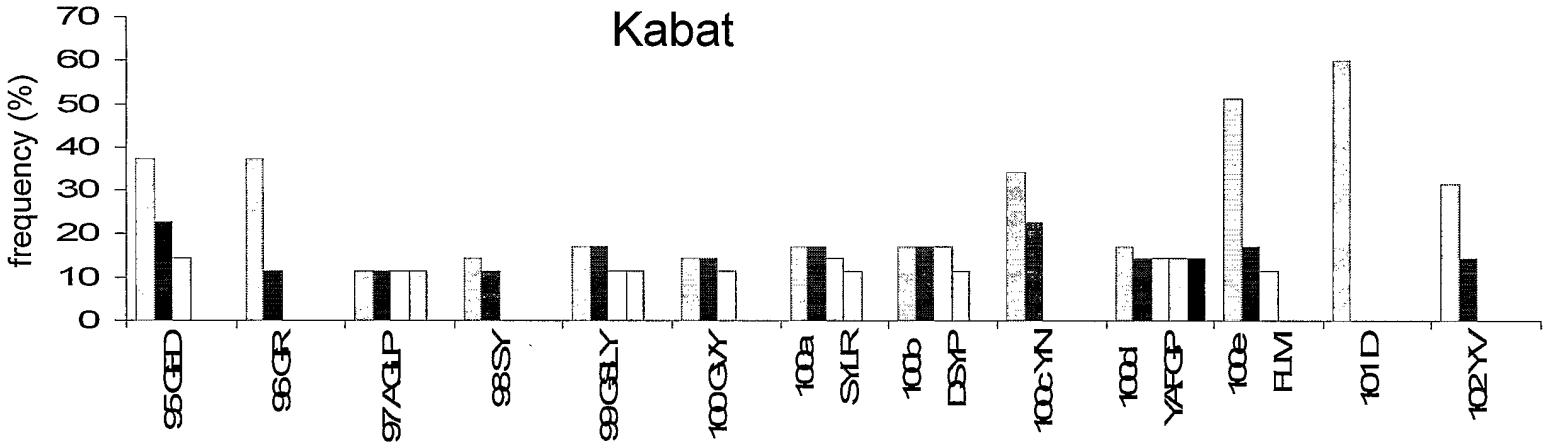


Fig. 11

- Position 100e and 101 are conserved
- Positions 95, 96, 97, 98, 99, 100, 100a, 100b, 100c, 100d and 102 are mutated.

XXXXXXXXXXFDX

- INVARIANT POSITIONS
- WTM IN UNIVERSAL LIBRARIES
- WTM IN AFFINITY MATURATION

V_H-CDR3 Sequence Matrices (Contact Definition)

	A	D	S	I	P	R	Y	H	N	M	K	C	E	F	G	L	Q	T	V
93	X																		
94						X													
95	X	X	X	X	X	X	X	X	X			X			X				X
96	X	X	X	X	X	X	X	X	X			X			X				X
97	X	X	X	X	X	X	X	X	X			X			X				X
98	X	X	X	X	X	X	X	X	X			X			X				X
99	X	X	X	X	X	X	X	X	X			X			X				X
100														X					
101		X																	

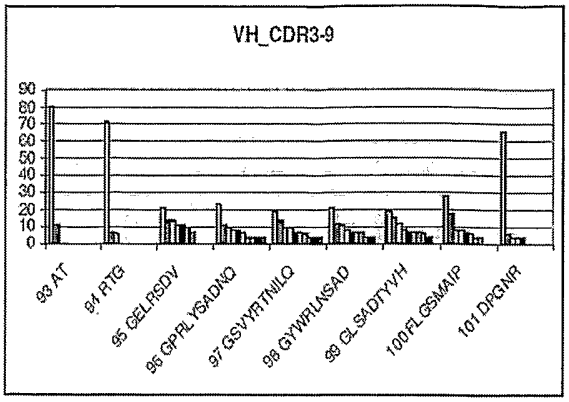
12 / 58

Fig. 12

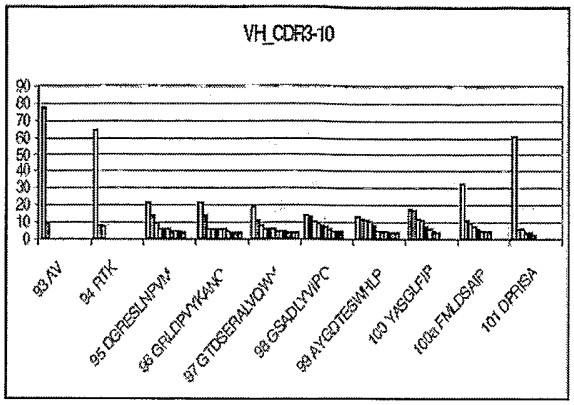
93	94	95	96	97	98	99	100	101
A	R	$X_{D/G}$	$X_{T/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	F	D

Fig. 12

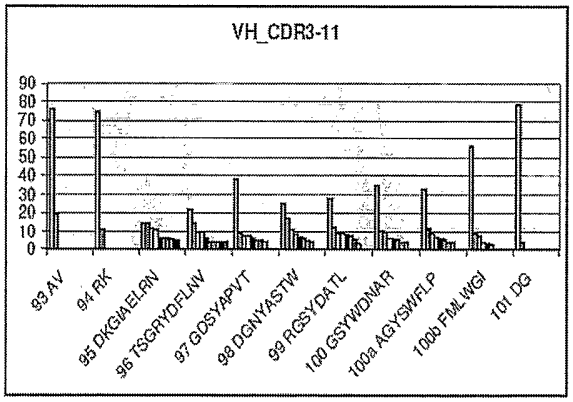
V_H-CDR3 Sequence Matrices (continued)



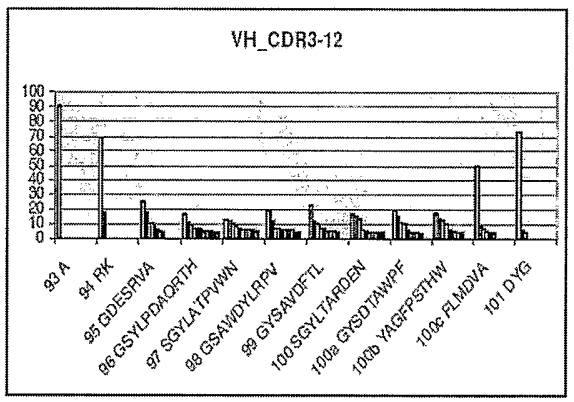
93	94	95	96	97	98	99	100	101
A	R	$X_{D/G}$	$X_{-R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	F	D



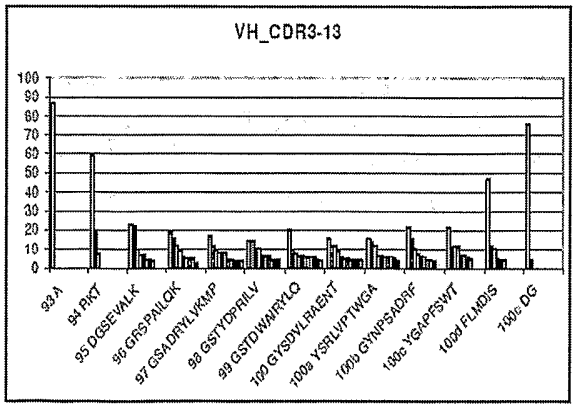
93	94	95	96	97	98	99	100	100A	101
A	R	$X_{D/G}$	$X_{-R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	X_Y	F	D



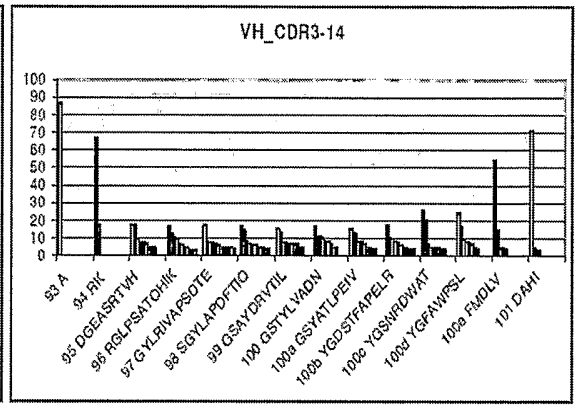
93	94	95	96	97	98	99	100	100A	100B	101
A	R	$X_{D/G}$	$X_{-R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	X_Y	X_Y	F	D



93	94	95	96	97	98	99	100	100A	100B	100C	101
A	R	$X_{D/G}$	$X_{R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	X_Y	X_Y	F	D



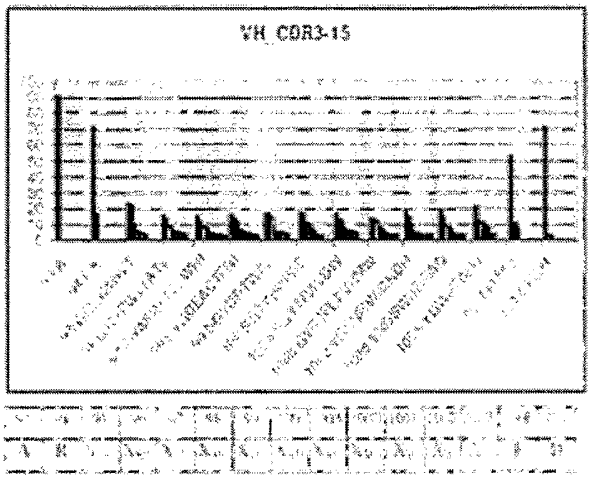
93	94	95	96	97	98	99	100	100A	100B	100C	100D	101
A	R	$X_{D/G}$	$X_{R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	X_Y	X_Y	X_Y	F	D

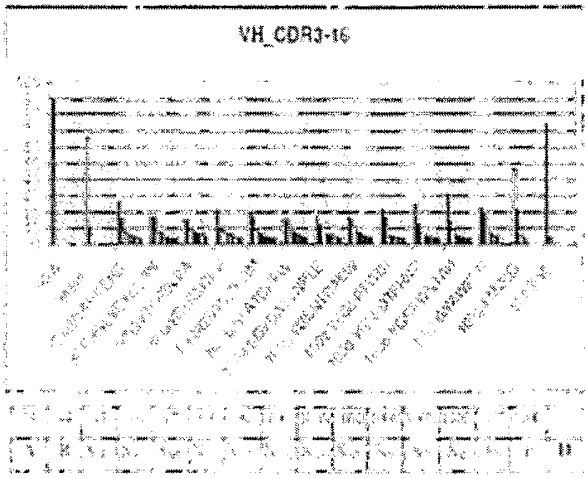


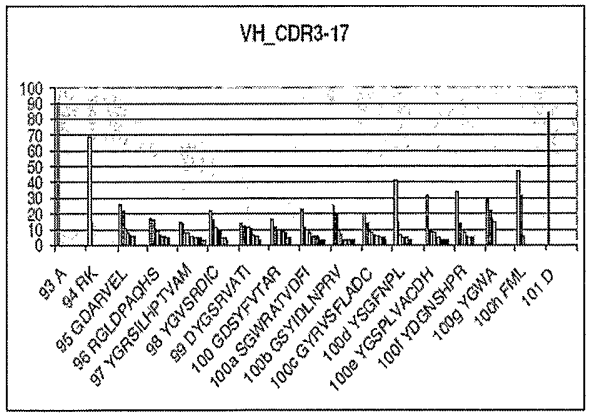
93	94	95	96	97	98	99	100	100A	100B	100C	100D	100E	101
A	R	$X_{D/G}$	$X_{R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{T/G}$	X_Y	X_Y	X_Y	F	D

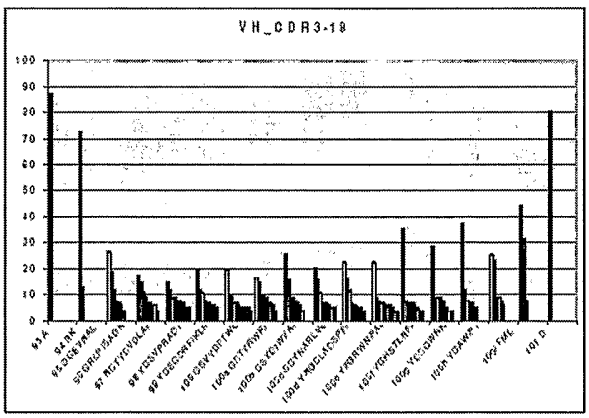
Fig. 12

V_H-CDR3 Sequence Matrices (continued)







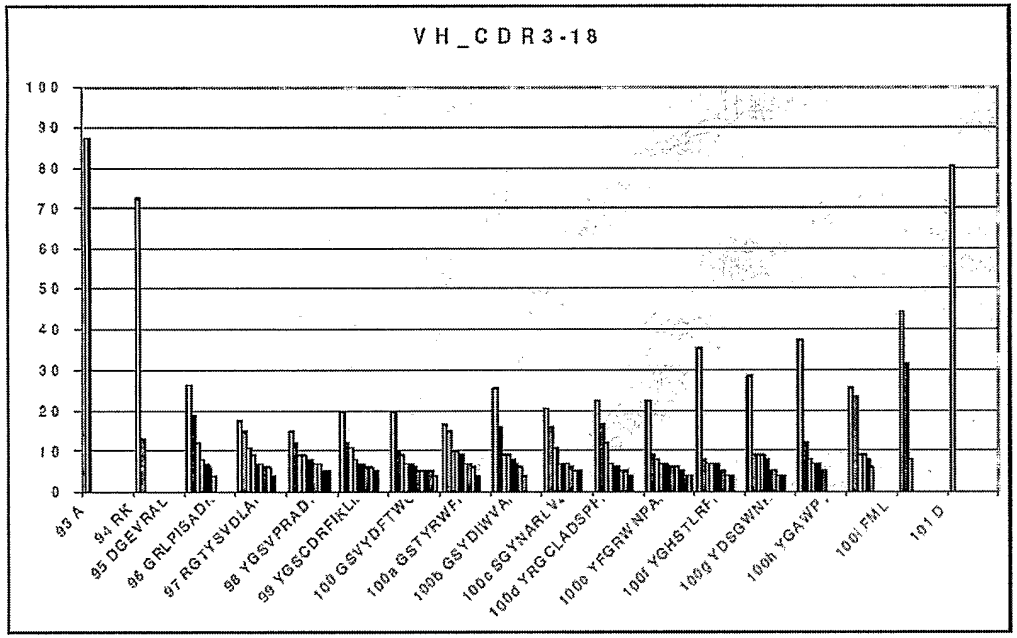


93	94	95	96	97	98	99	100	100A	100B	100C	100D	100E	100F	100G	100H	101
A	R	X_{DIC}	X_{RIL}	X_{SIG}	X_{SIC}	X_{SIG}	X_{SIG}	X_{SIG}	X_{SIC}	X_{SIG}	X_Y	X_Y	X_Y	X_Y	F	D

93	94	95	96	97	98	99	100	100A	100B	100C	100D	100E	100F	100G	100H	100I	101
A	R	X_{DIC}	X_{RIG}	X_{SIG}	X_{SIG}	X_{SIC}	X_{SIG}	X_{SIC}	X_{SIC}	X_{SIG}	X_Y	X_Y	X_Y	X_Y	X_Y	F	D

V_H-CDR3 Sequence Matrices (continued)

Fig. 12

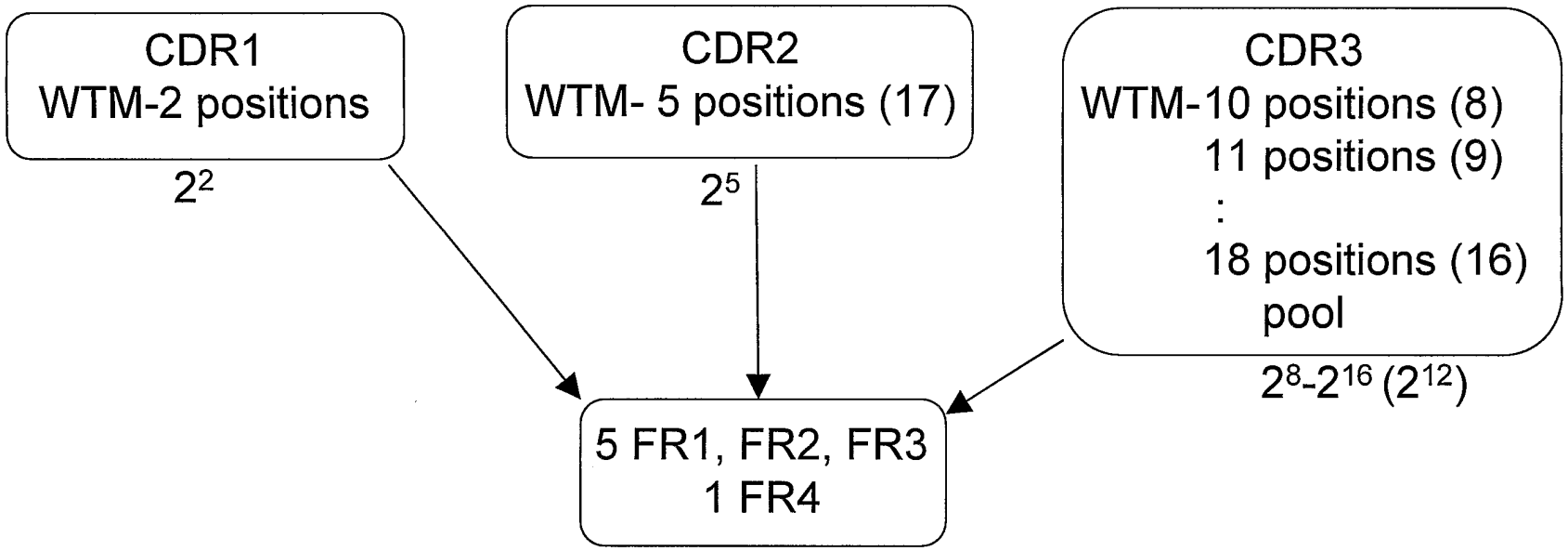


93	94	95	96	97	98	99	100	100A	100B	100C	100D	100E	100F	100G	100H	100I	101
A	R	$X_{D/G}$	$X_{R/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	$X_{S/G}$	X_Y	X_Y	X_Y	X_Y	X_Y	F	D

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
93	X																				
94						X															
95	X	X	X	X	X	X	X	X	X				X			X				X	
96	X	X	X	X	X	X	X	X	X				X			X				X	
97	X	X	X	X	X	X	X	X	X				X			X				X	X
98	X	X	X	X	X	X	X	X	X				X			X				X	X
99	X	X	X	X	X	X	X	X	X				X			X				X	X
100	X	X	X	X	X	X	X	X	X				X			X				X	X
100a	X	X	X	X	X	X	X	X	X				X			X				X	X
100b	X	X	X	X	X	X	X	X	X				X			X				X	X
100c	X	X	X	X	X	X	X	X	X				X			X				X	X
100d	X	X	X	X	X	X	X	X	X												
100e	X	X	X	X	X	X	X	X	X												
100f	X	X	X	X	X	X	X	X	X												
100g	X	X	X	X	X	X	X	X	X												
100h	X	X	X	X	X	X	X	X	X												
100i															X						
101		X																			

V_H library diversity

CDR diversity by WTM (low doping => 1 to 2 substitutions per molecule) with all amino acids except Cys.



V_H library diversity 10^5-10^7 (10^6)

- Combine V_H library with V_L library (V_κ : V_λ in 2:1)
- All oligos contain overlapping non-degenerate 20-mers on 5' and 3' ends

Fig. 13

Fig. 14

V_H LIBRARY CONSTRUCTION
FR1
CDR1
FR2
CDR2
rs
QVQLVQSGAEVKKPGSSVKVSCKASGGTF SSYAIS WVRQAPGQGLE WMGGIIPIFGTAN YAQKFQ
FR3
CDR3
FR4
linker
GRVTITADKSTSTAYMELSSLRSEDTAVYYC ARDGSGSGYAFD YWGQGTLVTVSS
Sense in blue
Antisense in red
Number of nondegenerate oligos needed
FR1: 4
FR2: 1
FR3: 5
FR4: 1
V_H
5 frameworks => 55 oligos
Number of degenerate oligos needed
CDR1: 2
CDR2: 2
CDR3: 10
X 9 (# of AA for WTM)

Fig. 15

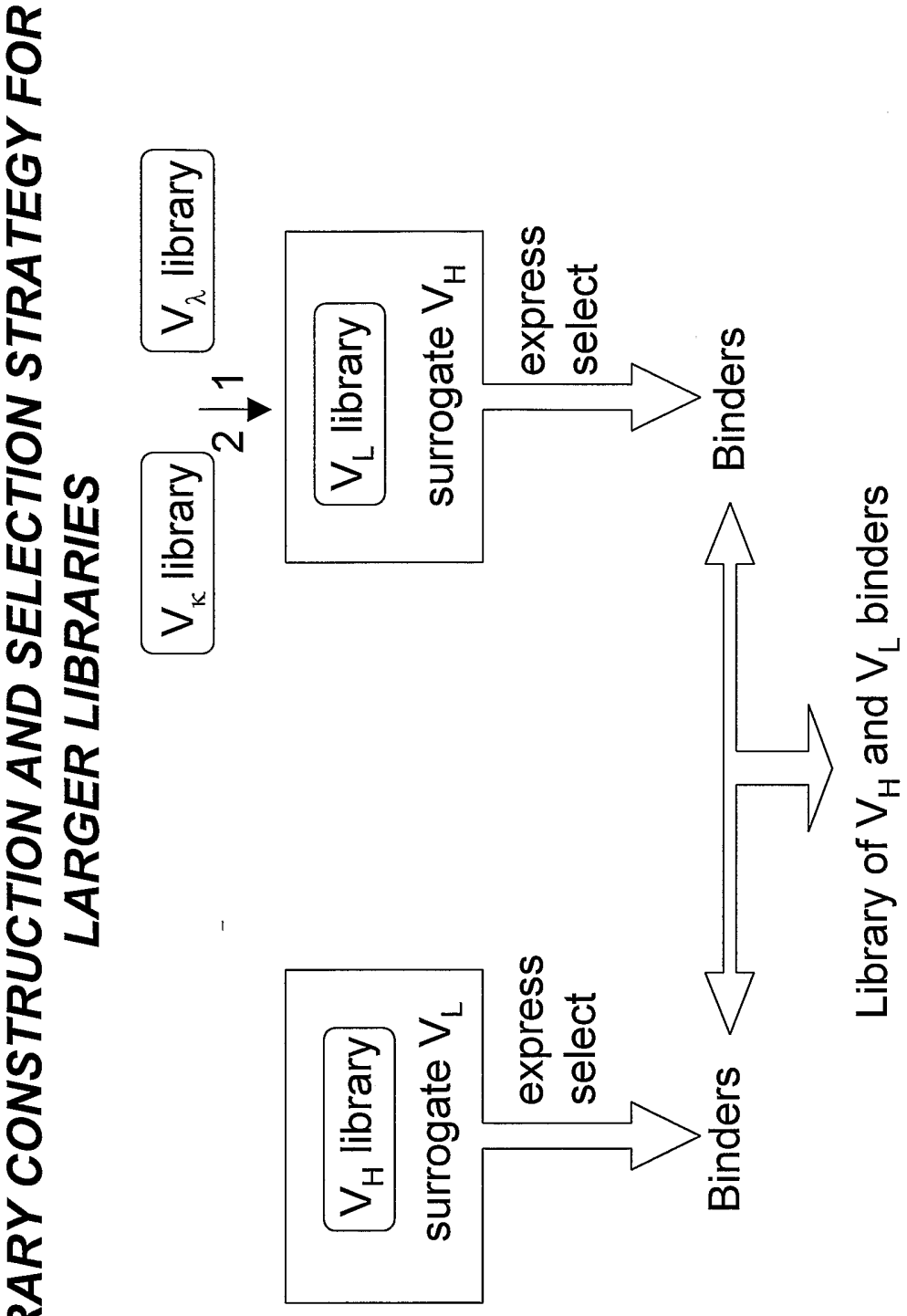

LIBRARY CONSTRUCTION AND SELECTION STRATEGY FOR LARGER LIBRARIES
V_κ library
V_λ library
2
1
V_L library
surrogate V_H
express
select
Binders
V_H library
surrogate V_L
express
select
Binders
Library of V_H and V_L binders

Fig. 16

HIGH FREQUENCY KAPPA (κ) AND LAMBDA (λ) LIGHT CHAIN FRAMEWORKS

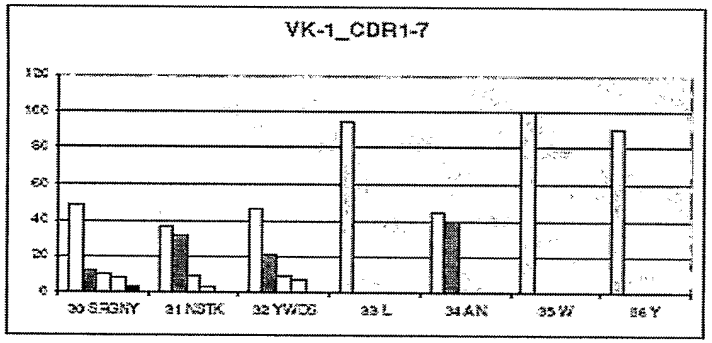
Vκ

	FR1			CDR1	FR2	CDR2		FR3				CDR3	FR4	
	1	2	3		4	5	6	7	8		9		0	
	12345678901234567890123456789			01a23456	789012345	6789012345	67890123456789012345678				90123456		7890123456a78	
I-L1	DIQMTQSPSSLSASVGDRVTITCRASQGI			SN-YLAWF	QQKPGKAPK	SLIYAASSLQ	SGVPSRFSGSGSGTDFTLTISSLQPEDFATYYC				QQYNSTPL		TFGQGTKVEIKRT	
III-A27	EIVLTQSPGTLSLSPGERATLSCRASQSV			SSSYLAWY	QQKPGQAPR	LLIYGASSRA	TGIPDRFSGSGSGTDFTLTISRLEPEDFAVYYC				QQYNSTPL		TFGQGTKVEIKRT	
III-L6	EIVLTQSPATLSLSPGERATLSCRASQSV			SS-YLAWY	QQKPGQAPR	LLIYDASNRA	TGIPARFSGSGSGTDFTLTISSLEPEDFAVYYC				QQYNSTPL		TFGQGTKVEIKRT	

Vλ

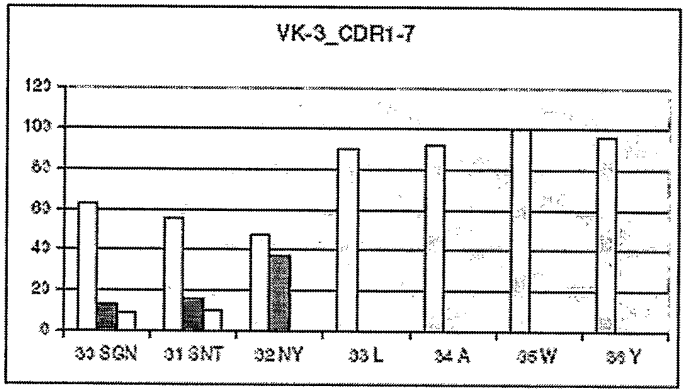
	FR1			CDR1	FR2	CDR2		FR3				CDR3	FR4	
	2		3		4	5	6	7	8		9		0	
	12345678912345678901234567890			01abc23456	789012345	6789012345	67890123456789012345678				9012345ab6		7890123456a78	
1b	QSVLTQPPSVSAAPGQKVTISCSGSSSN			IGNN-YVSWY	QQLPGTAPK	LLIYDNNKRP	SGIPDRFSGSKSGTSATLGITGLQTGDEADYYC				QSWDSSLNGV		VFGGGTKLTVLGQ	
2a2	QSALTQPASVSGSPGQSITISCTGTSSD			VGGYNYVSWY	QQHPGKAPK	LMIYEVSNRP	SGVSNRFSGSKSGNTASLTISGLQAEDEADYYC				QSWDSSLNGV		VFGGGTKLTVLGQ	
3l	SSELTQDPAVSVALGQTVRITCQGDSLR			SY---YASWY	QQKPGQAPV	LVIYGKNNRP	SGIPDRFSGSSSGNTASLTITGAQAEDEADYYC				QSWDSSLNGV		VFGGGTKLTVLGQ	
3r	SYELTQPPSVSVSPGQTASITCSGDKLG			DK---YACWY	QQKPGQSPV	LVIYQDSKRP	SGIPERFSGSNSGNTATLTISGTQAMDEADYYC				QSWDSSLNGV		VFGGGTKLTVLGQ	

$V_{L(Kappa)}$-CDR1 Sequence Matrices (Contact Definition)



30	31	32	33	34	35	36
S	X_S	X_Y	L	A–N	W	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30			X																		
31	X	X	X	X	X	X	X	X	X												
32	X	X	X	X	X	X	X	X	X												
33																	X				
34	X								X												
35																					X
36							X														



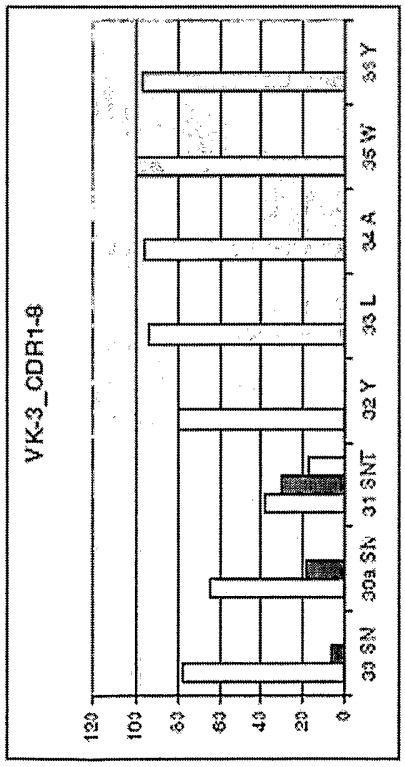
30	31	32	33	34	35	36
S	S	N–Y	L	A	W	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30			X																		
31			X																		
32							X		X												
33																	X				
34	X																				
35																					X
36							X														

Fig. 17

Fig. 17

$V_{L(Kappa)}$-CDR1 Sequence Matrices (continued)



30	30A	31	32	33	34	35	36
S	S-N	X_S	Y	L	A	W	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30			X																		
30A			X						X												
31	X	X	X	X	X	X	X	X	X												
32							X														
33																	X				
34	X																				
35																					X
36							X														

Fig. 18

VH3-23-VKIII-A27 Sequence SEQ ID NO: 1

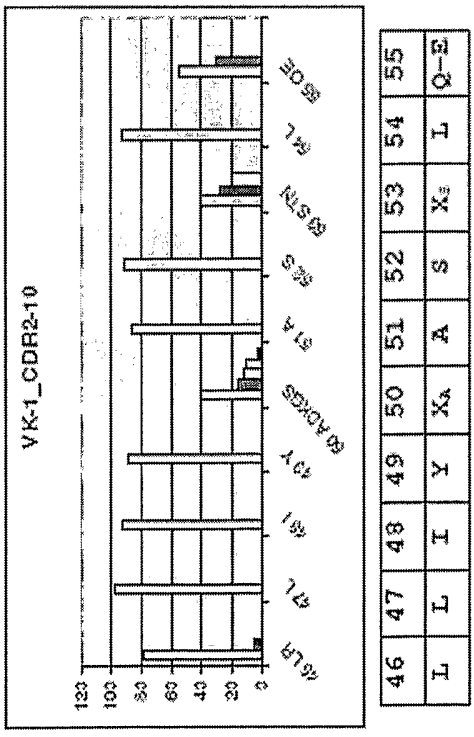
```
                                   CDR-H1
                                                                                ~~~~~~~~~~~~~~
     GluValGlnLeu LeuGluSer GlyGlyGly LeuValGlnPro GlyGlySer LeuArgLeu SerCysAlaAla SerGlyPhe ThrPheSer SerTyrAlaMet·
   1 GAAGTGCAGC TGTTGGAGTC TGGTGGAGGA TTGGTGCAGC CTGGCGGTAG CTTGCGCCTG TCTTGTGCTG CCAGTGGCTT TACCTTCTCT AGCTATGCTA
     CTTCACGTCG ACAACCTCAG ACCACCTCCT AACCACGTCG GACCGCCATC GAACGCGGAC AGAACACGAC GGTCACCGAA ATGGAAGAGA TCGATACGAT
     CDR-H1                                            CDR-H2
     ~~~~~                                   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
     ·MSerTrpVal ArgGlnAla ProGlyLysGly LeuGluTrp ValSerAla IleSerGlySer GlyGlySer ThrTyrTyr AlaAspSerVal LysGlyArg·
 101 TGAGTTGGGT TAGACAGGCT CCTGGCAAGG GTTTGGAATG GGTGTCTGCT ATCTCTGGCT CTGGCGGTAG CACCTACTAT GCAGATAGCG TCAAAGGCCG
     ACTCAACCCA ATCTGTCCGA GGACCGTTCC CAAACCTTAC CCACAGACGA TAGAGACCGA GACCGCCATC GTGGATGATA CGTCTATCGC AGTTTCCGGC
                                                                                                   CDR-H3
                                                                                                 ~~~~~~~~~~~~~
     ·PheThrIle SerArgAspAsn SerLysAsn ThrLeuTyr LeuGlnMetAsn SerLeuArg AlaGluAsp ThrAlaValTyr TyrCysAla LysAspGly
 201 CTTCACCATC AGCCGGGATA ACAGTAAAAA CACCCTGTAC TTGCAGATGA ACAGCCTGCG CGCCGAAGAT ACCGCTGTGT ATTACTGTGC TAAAGATGGT
     GAAGTGGTAG TCGGCCCTAT TGTCATTTTT GTGGGACATG AACGTCTACT TGTCGGACGC GCGGCTTCTA TGGCGACACA TAATGACACG ATTTCTACCA
              CDR-H3
     ~~~~~~~~~~~~~~~~~~~~~~~~~~
     SerGlySerGly TyrAlaPhe AspTyrTrp GlyGlnGlyThr LeuValThr ValSerSer GlyGlyGlyGly SerGlyGly GlyGlySer GlyGlyGlyGly·
 301 TCTGGTTCCG GCTACGCCTT CGATTACTGG GGTCAGGGCA CACTGGTTAC CGTCTCTAGC GGTGGAGGCG GTTCTGGTGG AGGCGGTTCG GGTGGCGGAG
     AGACCAAGGC CGATGCGGAA GCTAATGACC CCAGTCCCGT GTGACCAATG GCAGAGATCG CCACCTCCGC CAAGACCACC TCCGCCAAGC CCACCGCCTC
                                                                                                      CDR-L1
                                                                                                      ~~~~~~~~
     ·GSerGluIle ValLeuThr GlnSerProGly ThrLeuSer LeuSerPro GlyGluArgAla ThrLeuSer CysArgAla SerGlnSerVal SerSerSer·
 401 GTTCAGAAAT CGTGCTGACA CAGTCTCCAG GCACCTTGTC TCTGTCCCCA GGCGAACGCG CTACACTGTC CTGCAGAGCT TCTCAGTCCG TGTCTAGTTC
     CAAGTCTTTA GCACGACTGT GTCAGAGGTC CGTGGAACAG AGACAGGGGT CCGCTTGCGC GATGTGACAG GACGTCTCGA AGAGTCAGGC ACAGATCAAG
         CDR-L1                                          CDR-L2
     ~~~~~~~~~~~~~~~~                         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
     ·TyrLeuAla TrpTyrGlnGln LysProGly GlnAlaPro ArgLeuLeuIle TyrGlyAla SerSerArg AlaThrGlyIle ProAspArg PheSerGly
 501 CTATCTGGCC TGGTATCAAC AGAAACCTGG TCAGGCCCCT CGCTTGCTGA TCTACGGTGC TTCTAGCAGA GCCACAGGCA TCCCTGATAG ATTCTCTGGT
     GATAGACCGG ACCATAGTTG TCTTTGGACC AGTCCGGGGA GCGAACGACT AGATGCCACG AAGATCGTCT CGGTGTCCGT AGGGACTATC TAAGAGACCA
                                                                                        CDR-L3
                                                                                ~~~~~~~~~~~~~~~~~~~~~~~~~~
     SerGlySerGly ThrAspPhe ThrLeuThr IleSerArgLeu GluProGlu AspPheAla ValTyrTyrCys GlnGlnTyr AsnSerThr ProLeuThrPhe·
 601 AGCGGCTCTG GCACAGATTT CACACTGACT ATCTCCCGTT TGGAACCAGA AGATTTCGCC GTTTACTATT GCCAACAGTA CAACAGCACC CCATTGACAT
     TCGCCGAGAC CGTGTCTAAA GTGTGACTGA TAGAGGGCAA ACCTTGGTCT TCTAAAGCGG CAAATGATAA CGGTTGTCAT GTTGTCGTGG GGTAACTGTA
     ·PGlyGlnGly ThrLysVal GluIleLysArg Thr
 701 TCGGTCAGGG CACCAAAGTG GAAATCAAAA GAACC
     AGCCAGTCCC GTGGTTTCAC CTTTAGTTTT CTTGG
```

Fig. 19

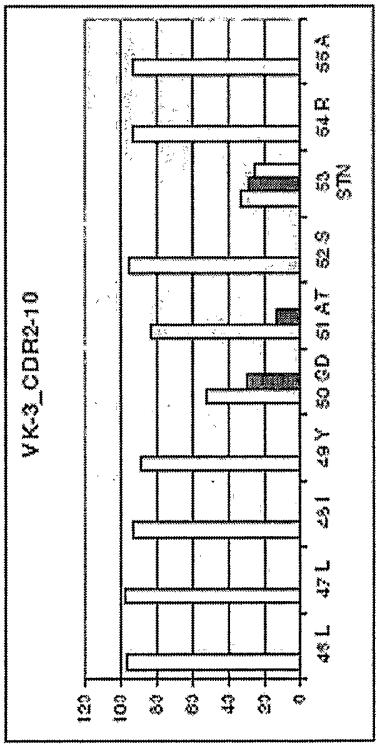
$V_{L(Kappa)}$-CDR2 Sequence Matrices (Contact Definition)

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
46																	X				
47																	X				
48				X			X														
49								X													
50	X	X	X	X	X	X	X	X	X												
51	X																				
52			X																		
53	X	X	X	X	X	X	X	X	X												
54																		X			
55														X					X		

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
46																	X				
47																	X				
48				X																	
49							X									X					
50		X																			
51	X																				
52			X																		
53	X	X	X		X	X		X	X												
54						X															
55	X																				

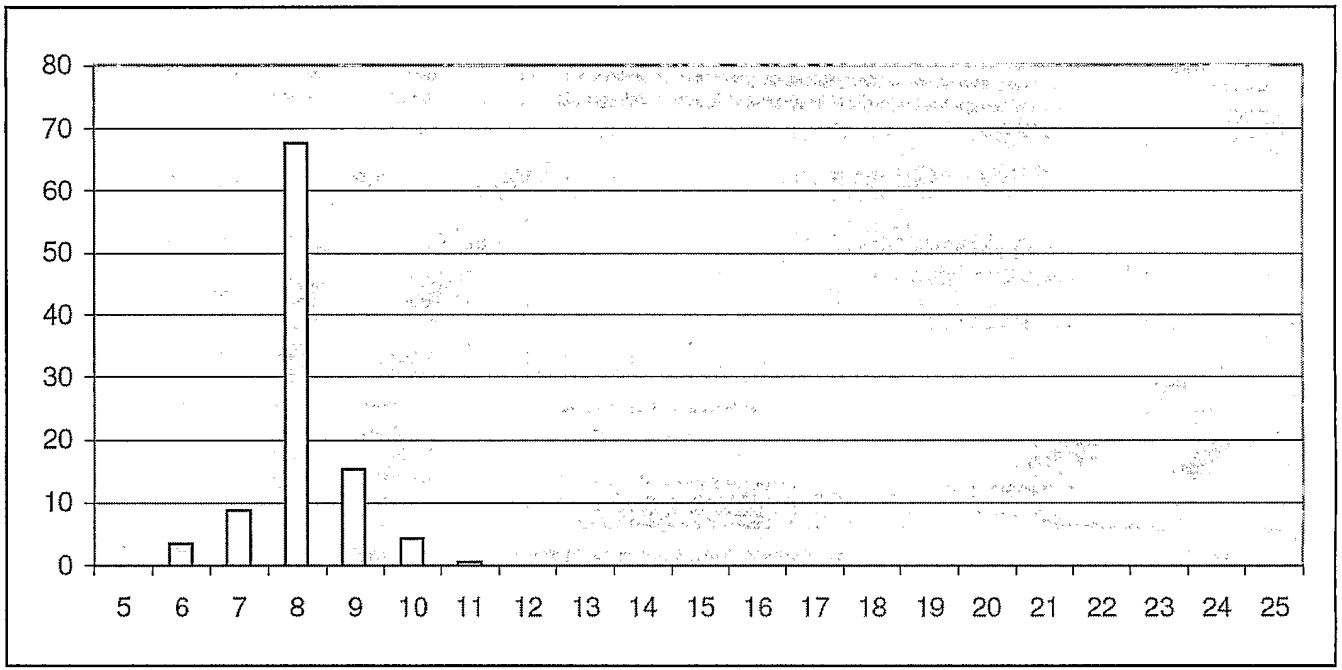


46	47	48	49	50	51	52	53	54	55
L	L	I	Y	X_A	A	S	X_S	L	Q–E



46	47	48	49	50	51	52	53	54	55
L	L	I	Y	G–D	A	S	X_S	R	A

V-kappa CDR3 length distribution



- The length variability of V_k CDR3 ranges from 6 to 11 residues (Contact Definition)
- However most (~80%) V_k CDR3s of anti-protein antibodies range from 8 to 9 residues
- Lengths 8 and 9 are synthesized separately and pooled before gene assembly in desired ratios

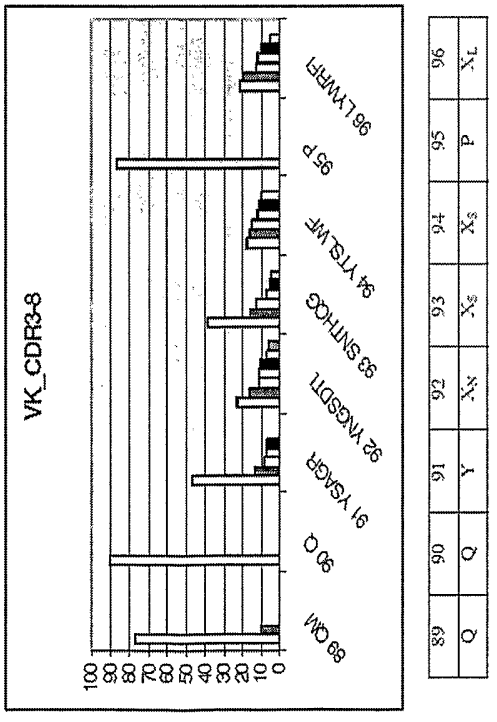
Fig. 20

Fig. 21

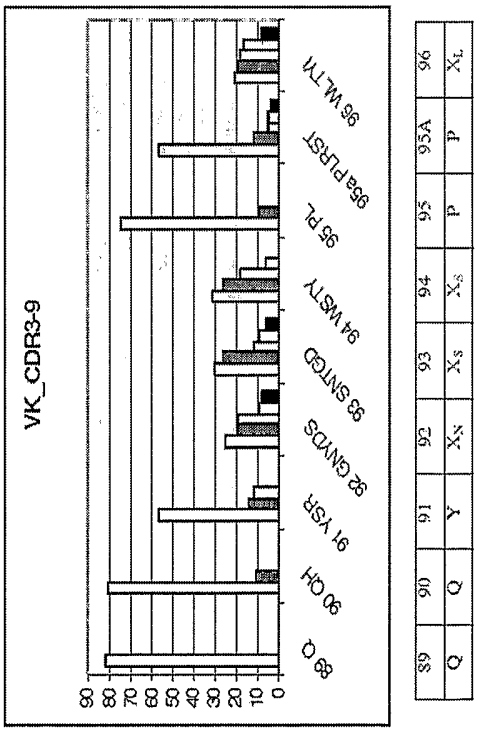
$V_{L(Kappa)}$-CDR3 Sequence Matrices (Contact Definition)

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																		X			
90																		X			
91							X														
92	X	X	X	X	X	X	X	X	X												
93	X	X	X	X	X	X	X	X	X												
94	X	X	X	X	X	X	X	X	X										X		
95					X																
96	X	X	X	X	X	X	X	X	X								X				

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																					
90																		X			
91							X											X			
92	X	X	X	X	X	X	X	X	X												
93	X	X	X	X	X	X	X	X	X												
94	X	X	X	X	X	X	X	X	X												
95					X																
95a						X															
96	X	X	X	X	X	X	X	X	X								X				

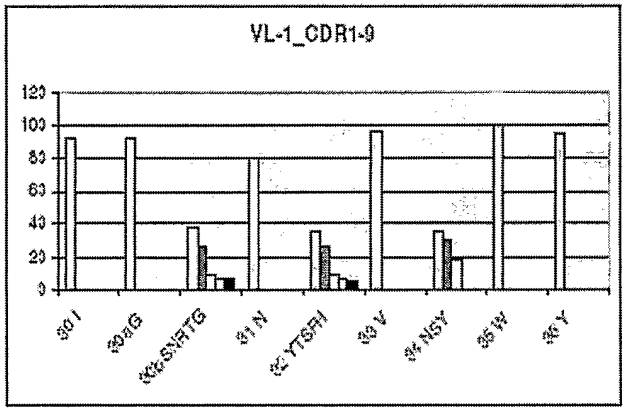


89	90	91	92	93	94	95	96
Q	Q	Y	X_N	X_S	X_S	P	X_L



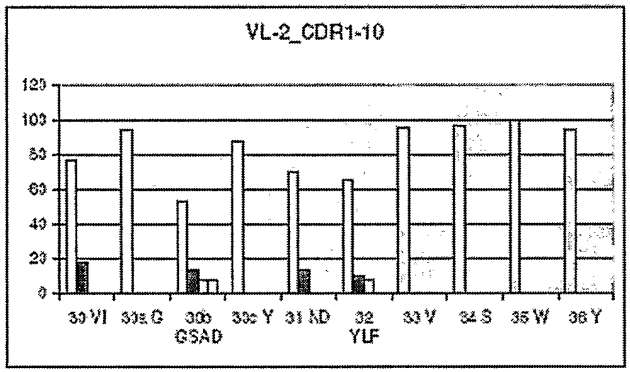
89	90	91	92	93	94	95	95A	96
Q	Q	Y	X_N	X_S	X_S	P	P	X_L

$V_{L(Lambda)}$-CDR1 Sequence Matrices (Contact Definition)



30	30A	30B	31	32	33	34	35	36
I	G	X_5	N	X_7	V	X_{11}	W	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30				X																	
30A																X					
30B	X	X	X	X	X	X	X	X	X												
31									X												
32	X	X	X	X	X	X	X	X	X												
33																				X	
34	X	X	X	X	X	X	X	X	X												
35																					X
36							X														



30	30A	30B	30C	31	32	33	34	35	36
V	G	X_5	Y	N	Y	V	S	W	Y

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30																				X	
30A																X					
30B	X	X	X	X	X	X	X	X	X							X					
30C							X														
31									X												
32							X														
33																				X	
34			X																		
35																					X
36							X														

Fig. 22

Fig. 22

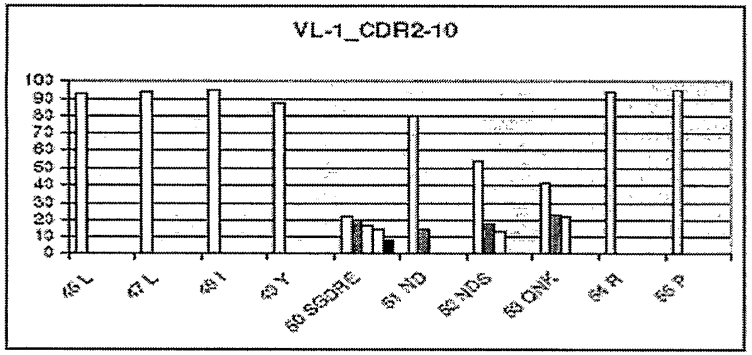
$V_{L(Lambda)}$-CDR1 Sequence Matrices (continued)

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
30	X	X	X	X	X	X	X	X	X												
31	X	X	X	X	X	X	X	X	X			X									
32	X	X	X	X	X	X	X	X	X											X	
33	X	X																			
34	X	X	X	X	X	X	X	X	X												X
35							X														
36																					

30	31	32	33	34	35	36
X_S	X_E	X_Y	A–V	X_H	W	Y

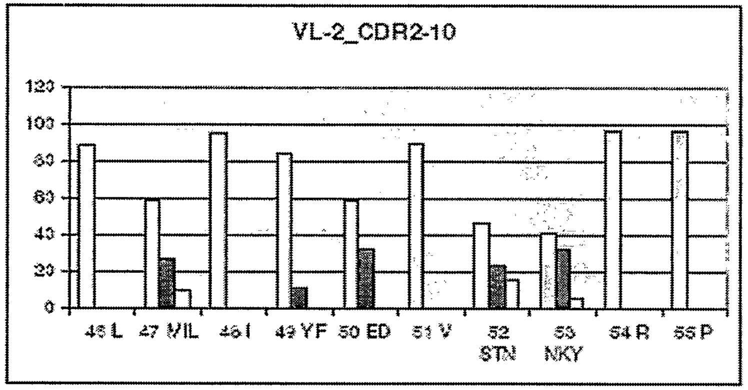
Fig. 23

$V_{L(Lambda)}$-CDR2 Sequence Matrices (Contact Definition)



46	47	48	49	50	51	52	53	54	55
L	L	I	Y	X_S	N	N-D	X_N	R	P

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
46																	X				
47																	X				
48				X																	
49							X														
50	X	X	X	X	X	X	X	X	X												
51									X												
52		X							X												
53	X	X	X	X	X	X	X	X	X												
54						X															
55					X																



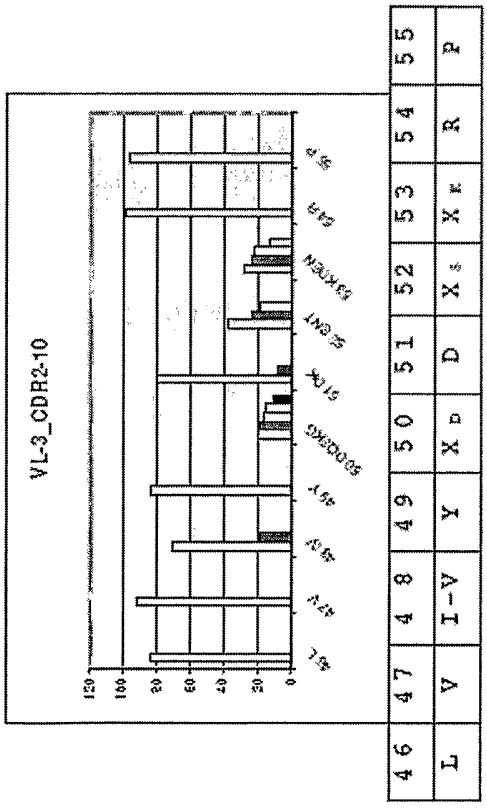
46	47	48	49	50	51	52	53	54	55
L	M-I	I	Y	E-D	V	X_S	X_N	R	P

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
46																	X				
47											X										
48				X																	
49							X														
50		X												X							
51																				X	
52	X	X	X	X	X	X	X	X	X										X		
53	X	X	X	X	X	X	X	X	X												
54								X													
55					X																

Fig. 23

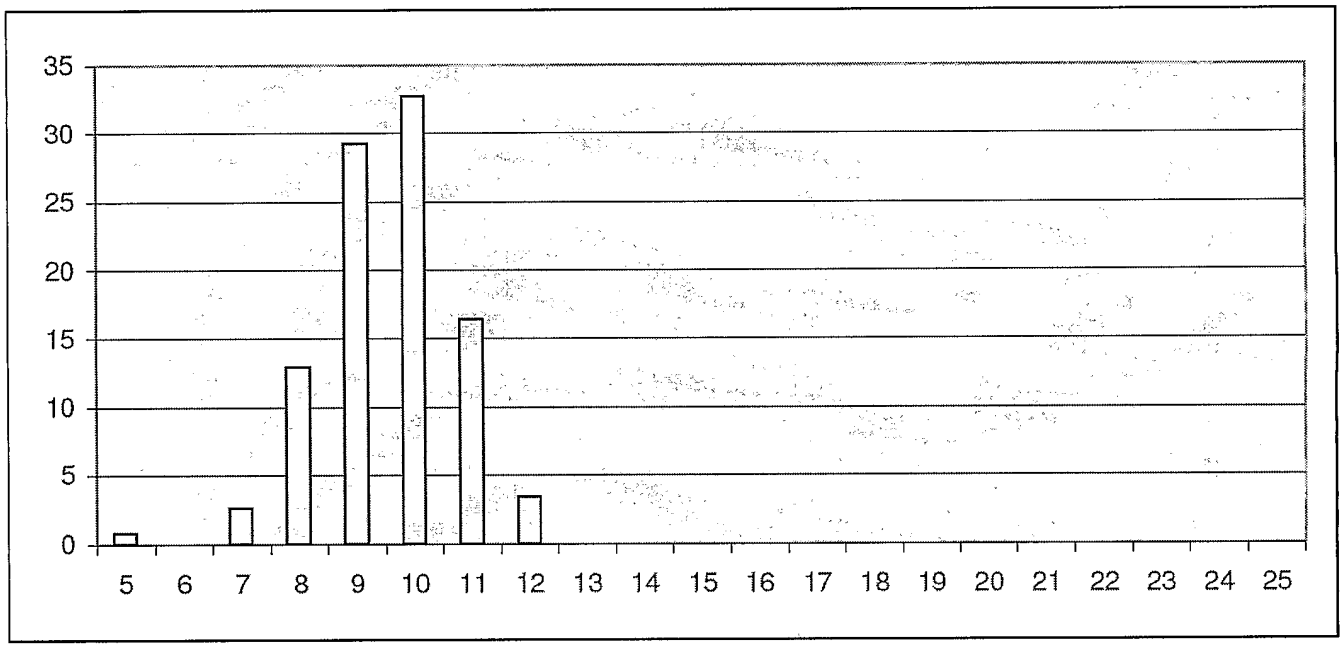
$V_{L(Lambda)}$-CDR2 Sequence Matrices (continued)

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
46																	X				
47																				X	
48				X																X	
49							X														
50	X	X	X	X	X	X	X	X	X											V	
51		X																			
52	X	X	X	X	X	X	X	X	X												
53	X	X	X	X	X	X	X	X	X			X									
54						X															
55					X																



46	47	48	49	50	51	52	53	54	55
L	V	I-V	Y	X D	D	X S	X K	R	P

V-lambda CDR3 length distribution



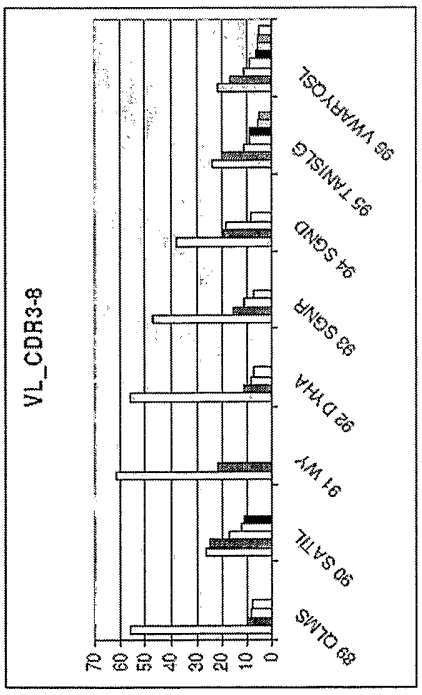
- The length variability of Vλ CDR3 ranges from 5 to 12 residues (Contact Definition)
- However most (~90%) V λ CDR3s of anti-protein antibodies range from 8 to 11 residues
- Lengths 8, 9, 10 and 11 are synthesized separately and pooled before gene assembly in desired ratios

Fig. 25

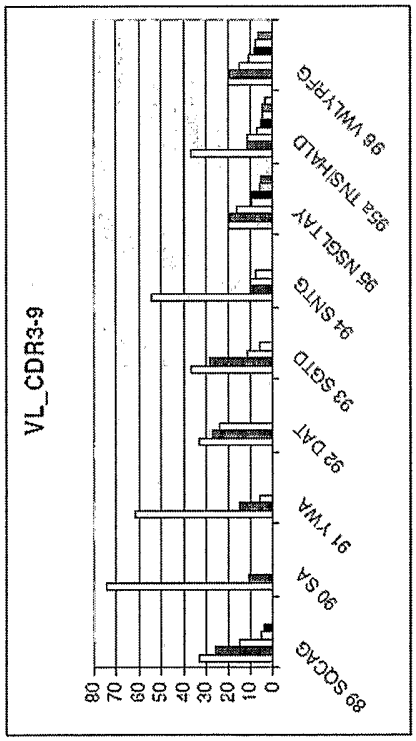
$V_{L(Lambda)}$-CDR3 Sequence Matrices (Contact Definition)

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																					
90	X		X																		
91																					X
92		X																			
93	X	X	X	X	X	X	X	X	X												
94			X																		
95	X	X	X	X	X	X	X	X	X												
96	X	X	X	X	X	X	X	X	X											X	

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																					
90	X		X																		
91							X														
92	X	X																			
93	X	X	X	X	X	X	X	X	X												
94		X																			
95	X	X	X	X	X	X	X	X	X												
95a	X	X	X	X	X	X	X	X	X											X	
96	X	X	X	X	X	X	X	X	X												X

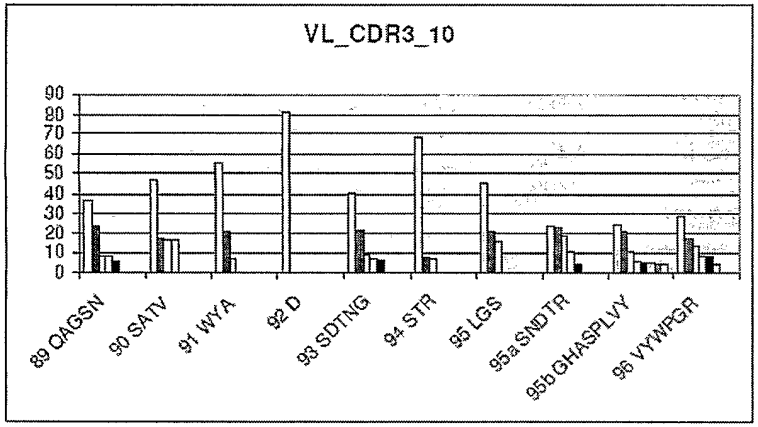


89	90	91	92	93	94	95	96
Q	S-A	W	D	X_S	S	X_N	X_Y



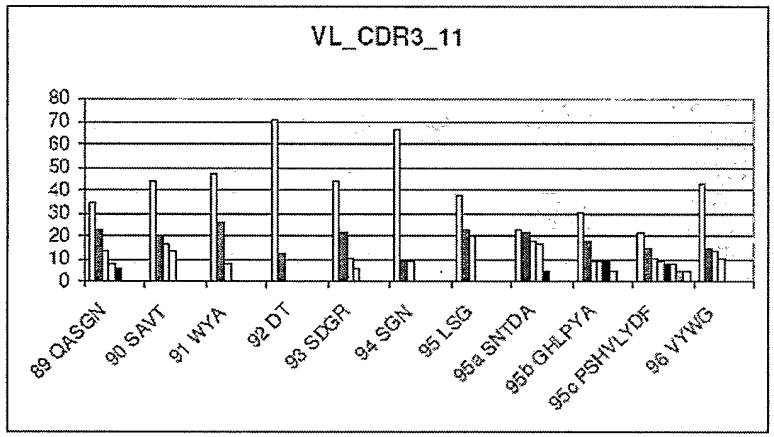
89	90	91	92	93	94	95	95A	96
Q	S-A	Y	D-A	X_S	S	X_N	X_T	X_Y

$V_{L(Lambda)}$-CDR3 Sequence Matrices (continued)



89	90	91	92	93	94	95	95A	95B	96
Q	S-A	W	D	X_S	S	L-S	X_N	X_G	X_V

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																		X			
90	X		X																		
91																					X
92		X																			
93	X	X	X	X	X	X	X	X	X												
94			X																		
95			X														X				
95a	X	X	X	X	X	X	X	X	X										X		
95b	X	X	X	X	X	X	X	X	X							X					
96	X	X	X	X	X	X	X	X	X											X	



89	90	91	92	93	94	95	95A	95B	995C	96
Q	S-A	W	D	X_S	S	L-S	X_N	X_G	X_P	X_V

	A	D	S	I	P	R	Y	H	N		M	K	C	E	F	G	L	Q	T	V	W
89																		X			
90	X		X																		
91																					X
92		X																			
93	X	X	X	X	X	X	X	X	X												
94			X																		
95			X														X				
95a	X	X	X	X	X	X	X	X	X										X		
95b	X	X	X	X	X	X	X	X	X							X					
95c	X	X	X	X	X	X	X	X	X											X	
96	X	X	X	X	X	X	X	X	X											X	

Fig. 25

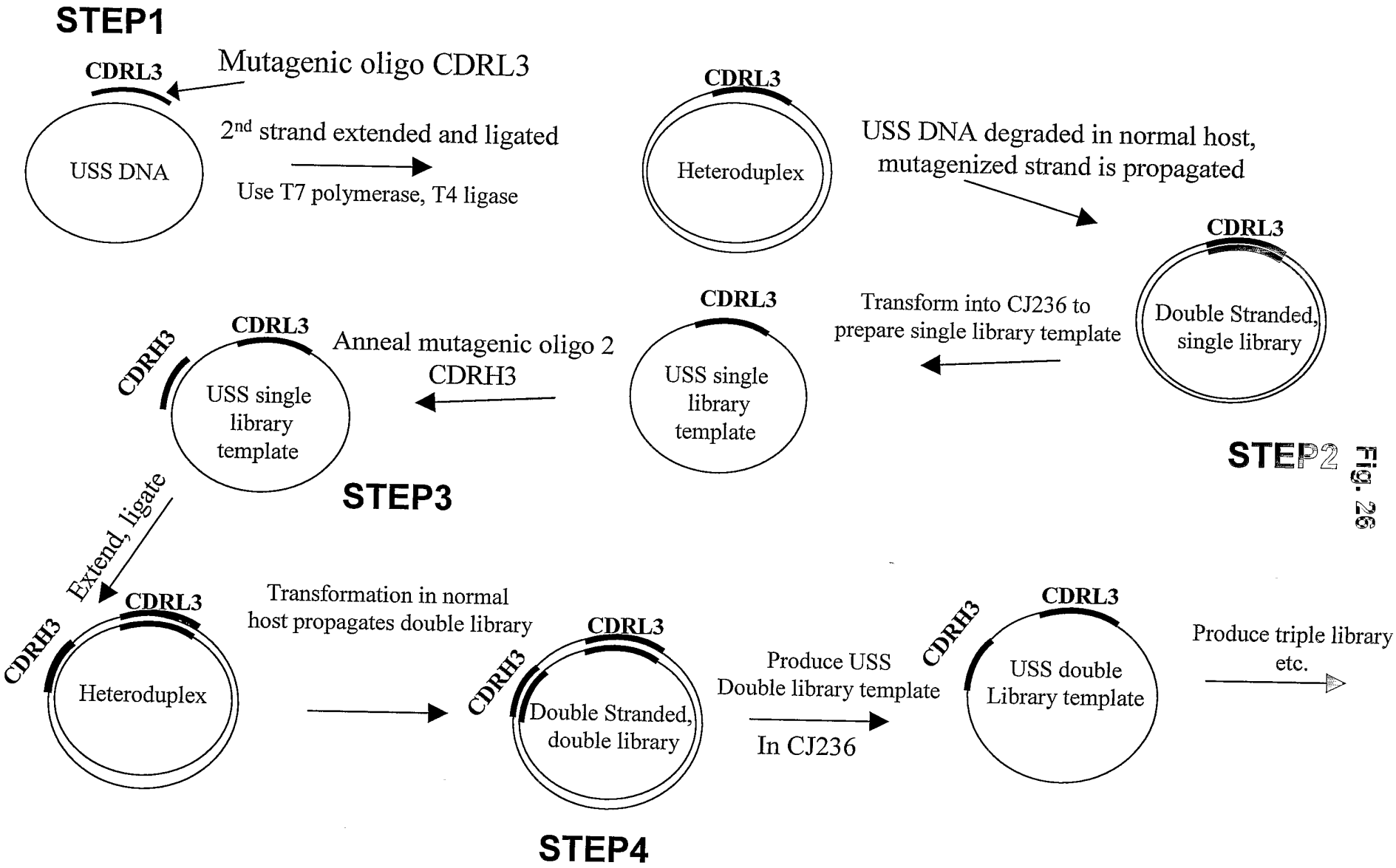
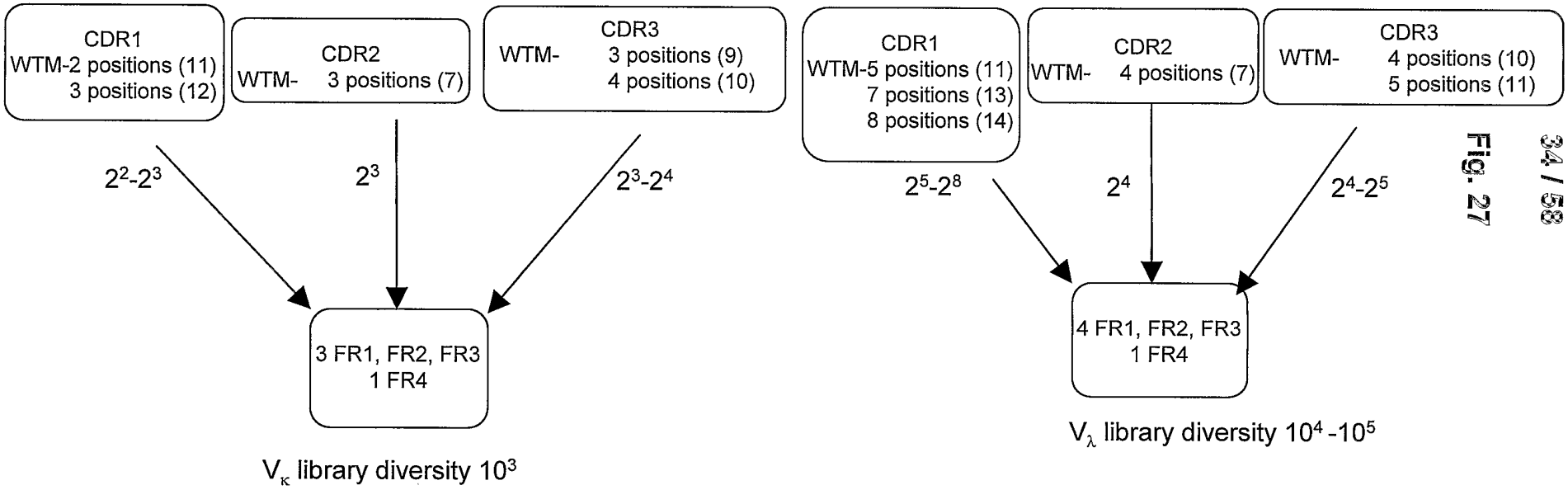

STEP1
CDRL3
Mutagenic oligo CDRL3
USS DNA
2nd strand extended and ligated
Use T7 polymerase, T4 ligase
Heteroduplex
USS DNA degraded in normal host, mutagenized strand is propagated
CDRL3
Double Stranded, single library
Transform into CJ236 to prepare single library template
CDRL3
USS single library template
Anneal mutagenic oligo 2 CDRH3
CDRH3
CDRL3
USS single library template
STEP2
STEP3
Extend, ligate
CDRH3
CDRL3
Heteroduplex
Transformation in normal host propagates double library
CDRL3
CDRH3
Double Stranded, double library
Produce USS Double library template
In CJ236
CDRH3
CDRL3
USS double Library template
Produce triple library etc.
STEP4

Fig. 2G

V_L library diversity

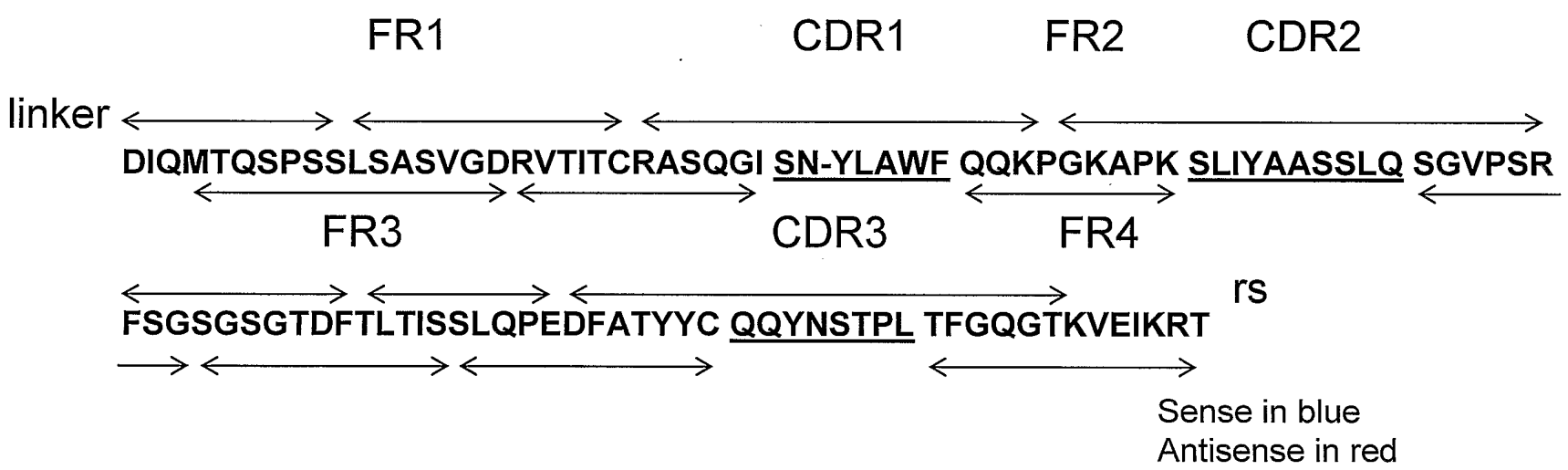
CDR diversity by WTM (low doping => 1 to 2 substitutions per molecule) with all amino acids except Cys.



- Combine V_L library with VH library (~10^{11})
- All oligos contain overlapping non-degenerate 20-mers on 5' and 3' ends

Fig. 27

$V_{L\ (Kappa\ \&\ Lambda)}$ LIBRARY CONSTRUCTION



Number of nondegenerate oligos needed

FR1:	4
FR2:	1
FR3:	5
FR4:	1

V_{kappa}: 3 frameworks => 33 oligos

V_{lambda}: 3 frameworks => 33 oligos

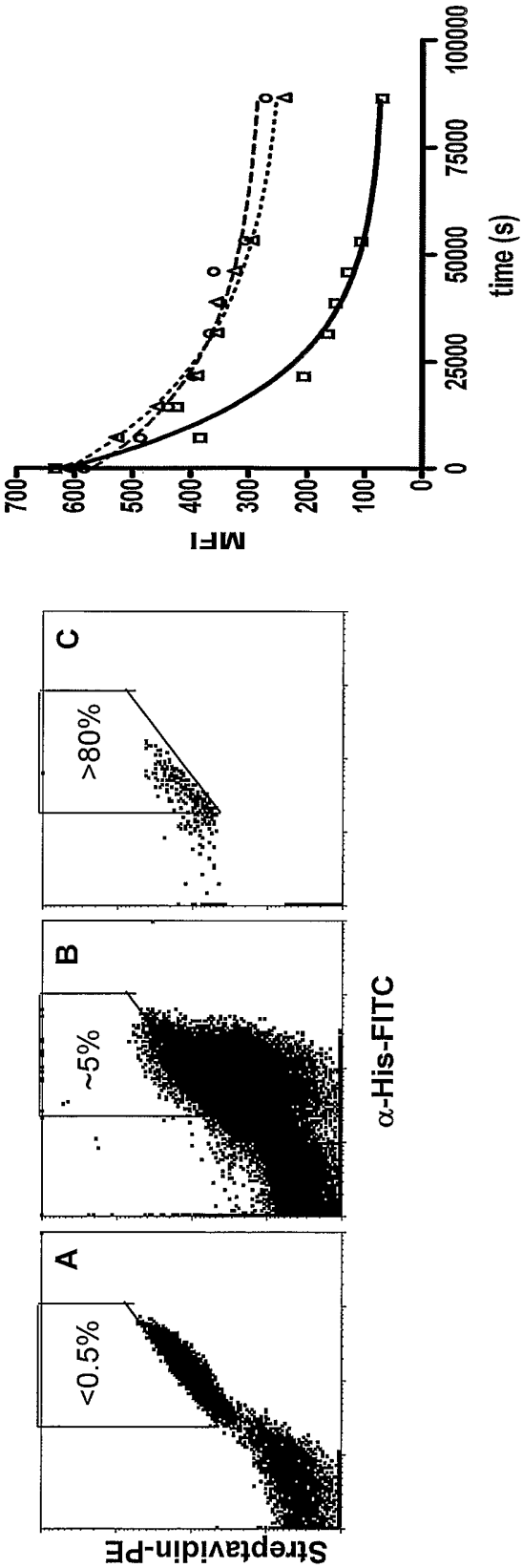
Number of degenerate oligos needed

	V_{kappa}	V_{lambda}
CDR1:	3	3
CDR2:	2	3
CDR3:	2	4

X 9 (# of AA for WTM)

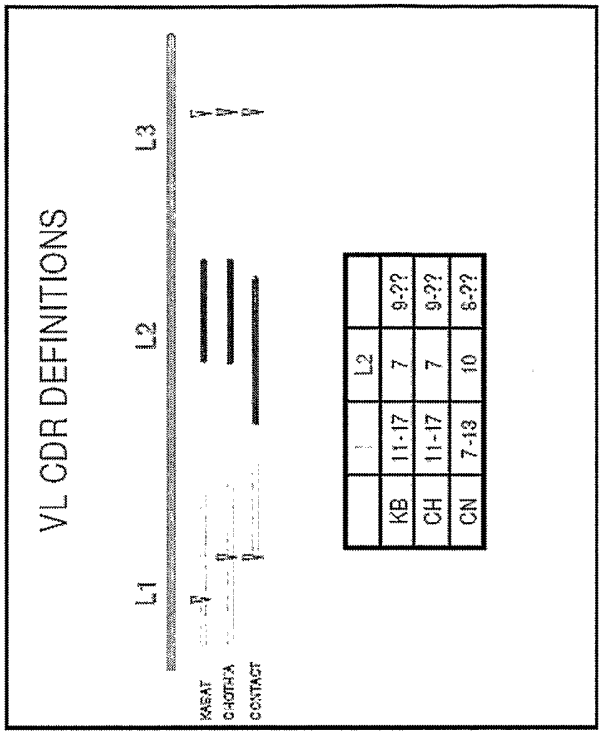
Fig. 28

Fig. 29



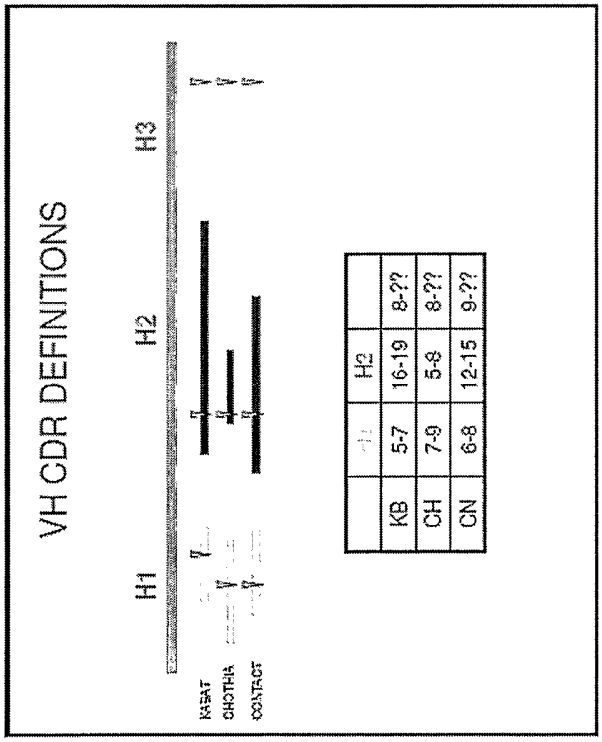
	CDR1			CDR2								CDR3								CDR4								CDR5						CDR6									
wt	D	M	H	Y	A	W	T	N	S	V	H	I	Y	A	D	D	S	G	E	Y	S	Y	S	L	T	L	D	S	A	S	V	A	R	N	L	A	I	R	S	Y	Q	G	S	Q	L	T	S	A	R	Q	Y	A	R	N	T	Y	P
		H	L		S	Y															Q	K					Q	Q					H				L			S	K	P	L	F	S	Y				P	K	D			
		Q			H																S	Q															K																
			Y																																			Y																
																																						K																

Fig. 30



	L1	L2	
KB	11-17	7	9-??
CH	11-17	7	9-??
CN	7-18	10	8-??

(B)



	H1	H2	
KB	5-7	16-19	8-??
CH	7-9	5-8	8-??
CN	6-8	12-15	9-??

(A)

Fig. 31

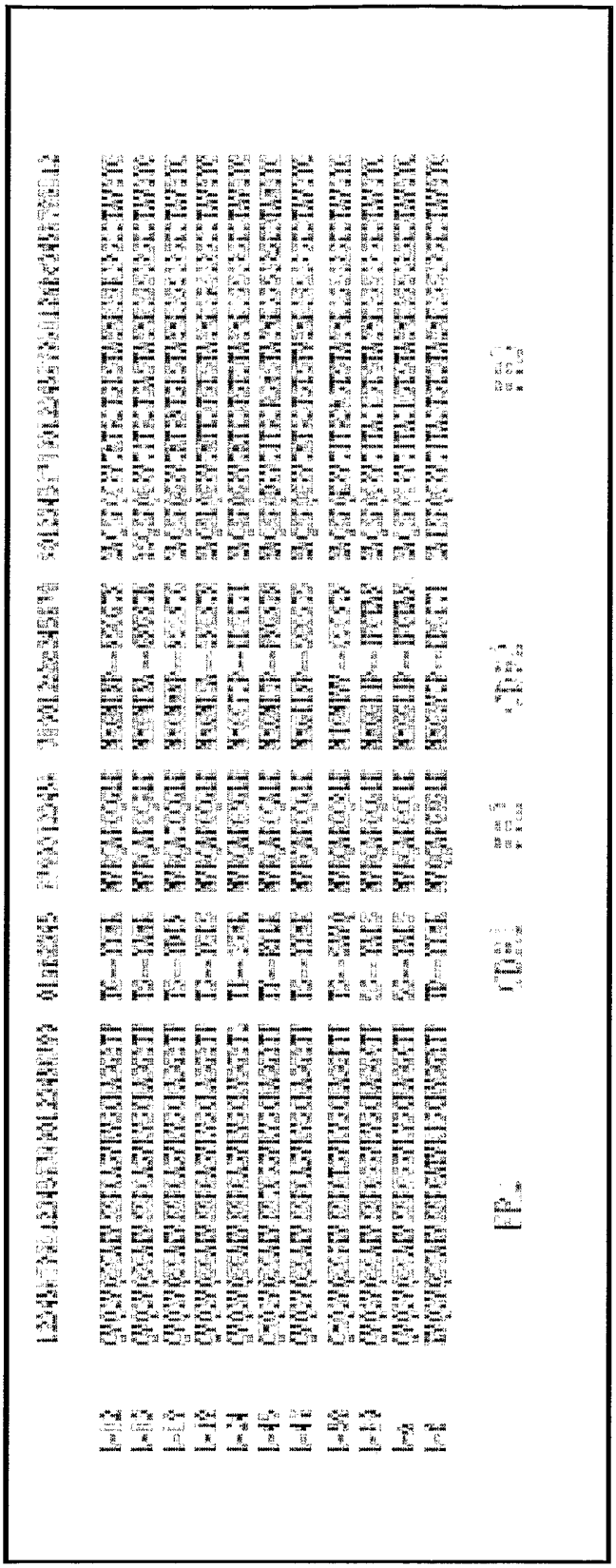


Fig. 32

VB_VH_FR123_CN.FASTA

>1-02
QVQLVQSGAEVKKPGASVKVSCKASGYTFxWVRQAPGQGLExYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYC
>1-03
QVQLVQSGAEVKKPGASVKVSCKASGYTFxWVRQAPGQRLExYSQKFQGRVTITRDTSASTAYMELSSLRSEDTAVYYC
>1-08
QVQLVQSGAEVKKPGASVKVSCKASGYTFxWVRQATGQGLExYAQKFQGRVTMTRNTSISTAYMELSSLRSEDTAVYYC
>1-18
QVQLVQSGAEVKKPGASVKVSCKASGYTFxWVRQAPGQGLExYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYC
>1-24
QVQLVQSGAEVKKPGASVKVSCKVSGYTLxWVRQAPGKGLExYAQKFQGRVTMTEDTSTDTAYMELSSLRSEDTAVYYC

Fig. 33

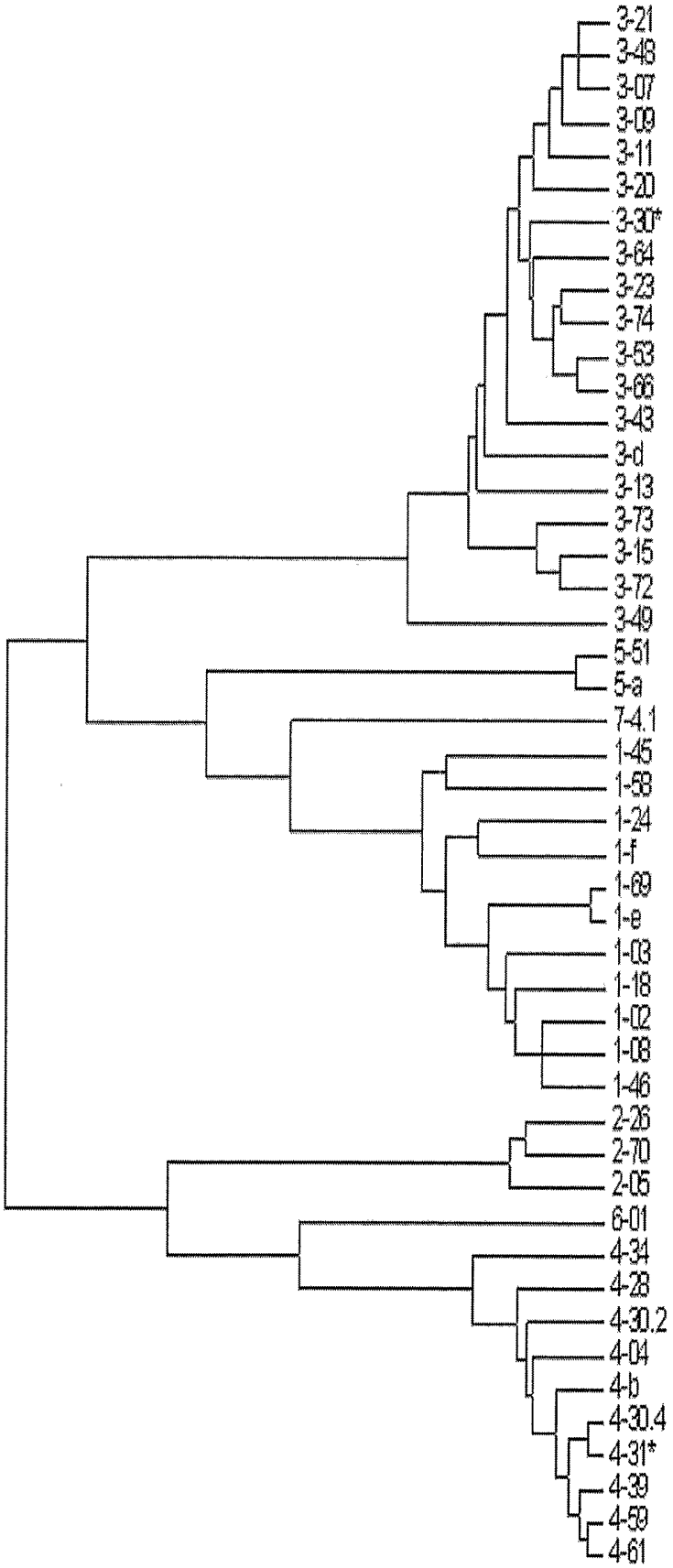

3-21
3-48
3-07
3-09
3-11
3-20
3-30*
3-64
3-23
3-74
3-53
3-66
3-43
3-d
3-13
3-73
3-15
3-72
3-49
5-51
5-a
7-4.1
1-45
1-58
1-24
1-f
1-69
1-e
1-03
1-18
1-02
1-08
1-46
2-26
2-70
2-05
6-01
4-34
4-28
4-30.2
4-04
4-b
4-30.4
4-31*
4-39
4-59
4-61

Fig. 34

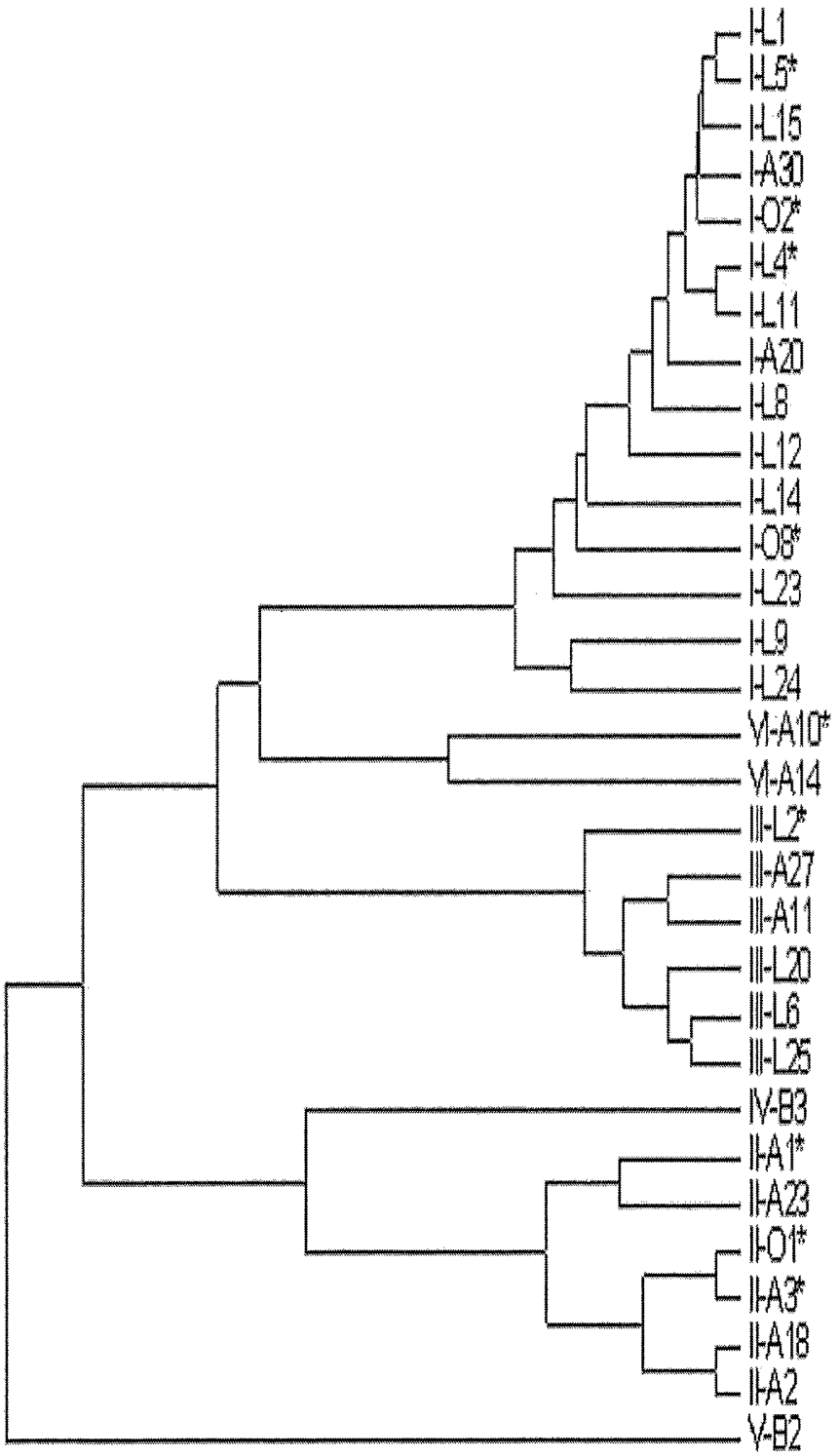

I-L1
I-L5*
I-L15
I-A30
I-O2*
I-L4*
I-L11
I-A20
I-L8
I-L12
I-L14
I-O8*
I-L23
I-L9
I-L24
VI-A10*
VI-A14
III-L2*
III-A27
III-A11
III-L20
III-L6
III-L25
IV-B3
II-A1*
II-A23
II-O1*
II-A3*
II-A18
II-A2
V-B2

Fig. 35

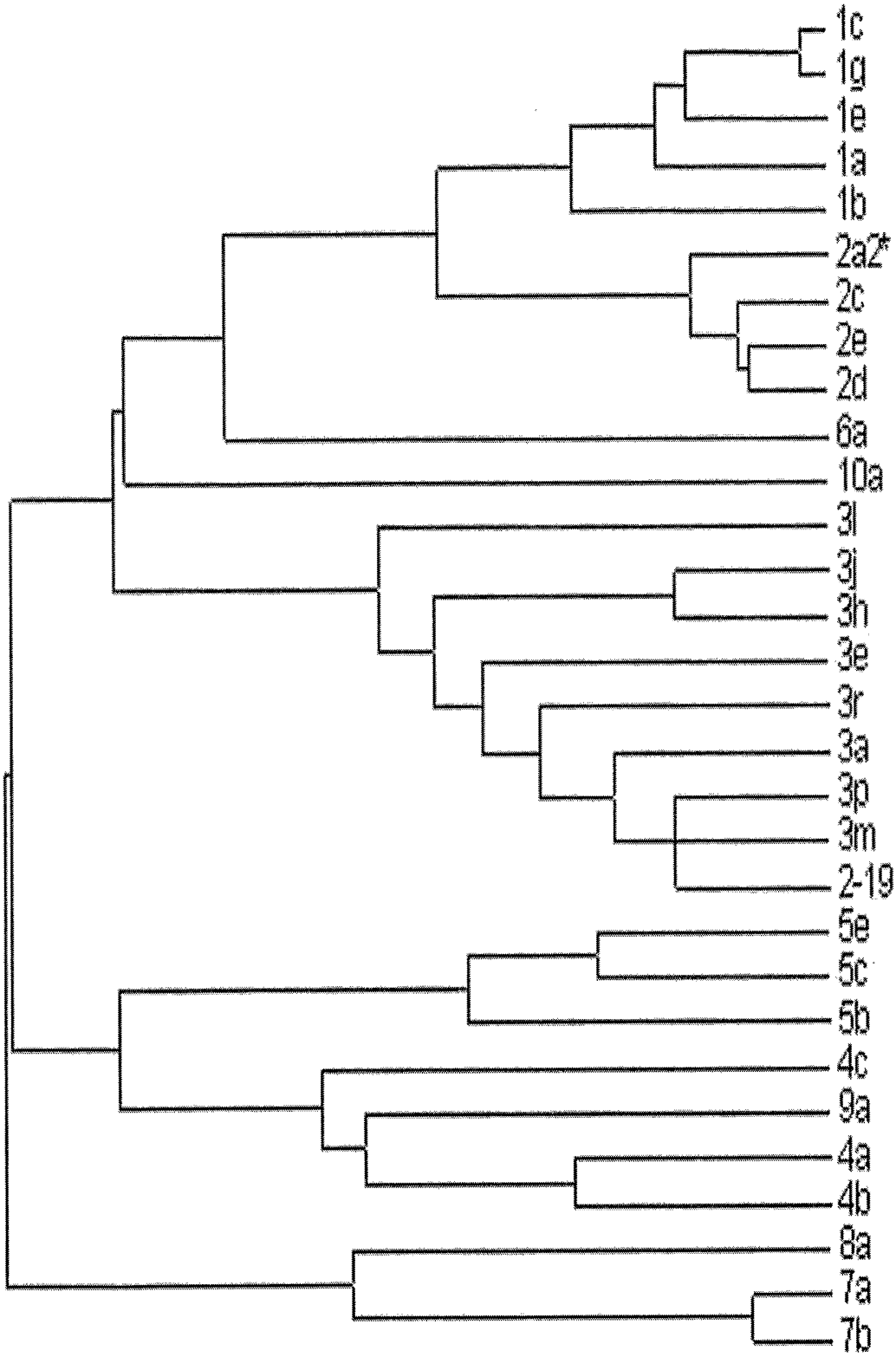

1c
1g
1e
1a
1b
2a2*
2c
2e
2d
6a
10a
3l
3j
3h
3e
3r
3a
3p
3m
2-19
5e
5c
5b
4c
9a
4a
4b
8a
7a
7b

Fig. 36

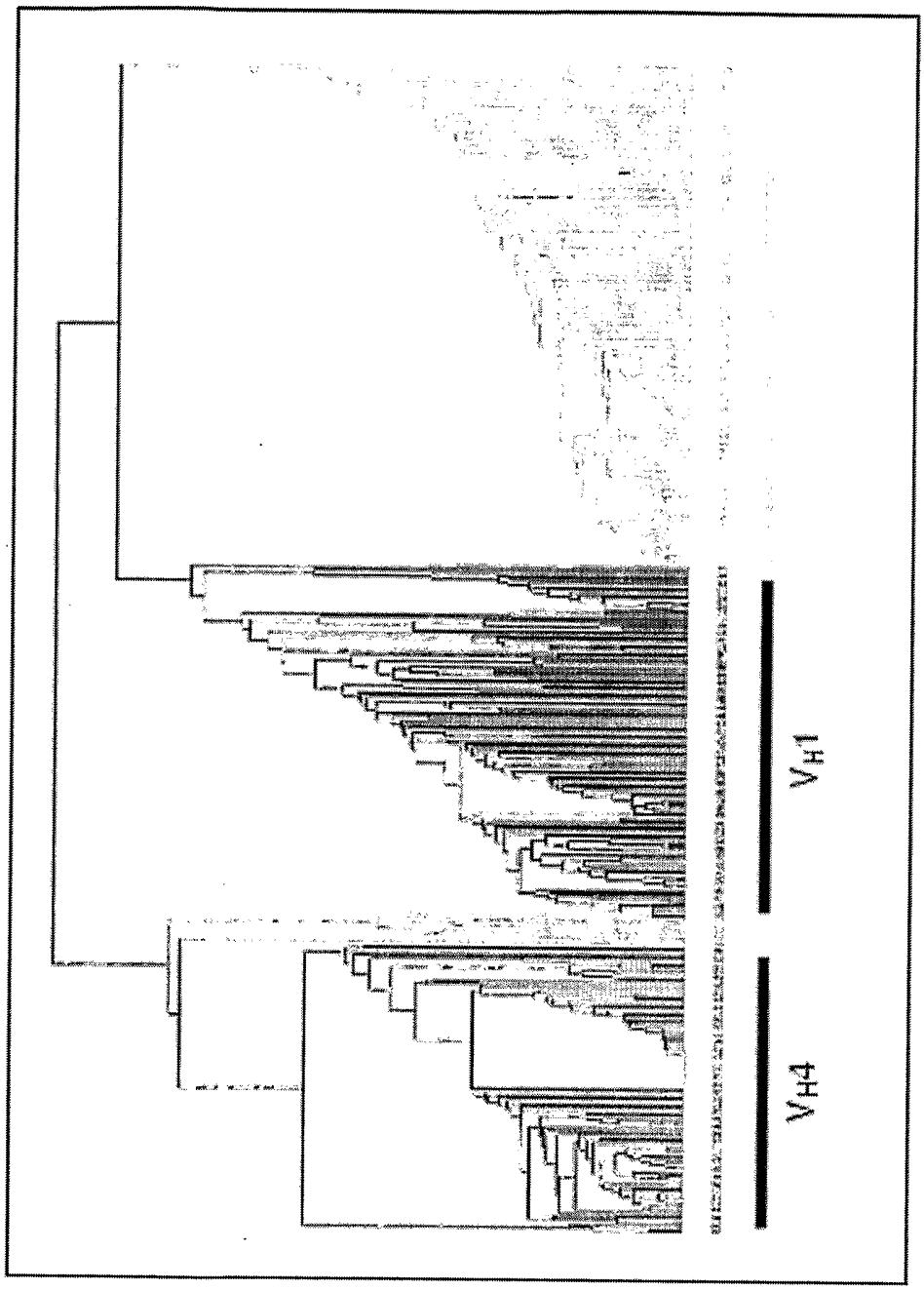


Fig. 37

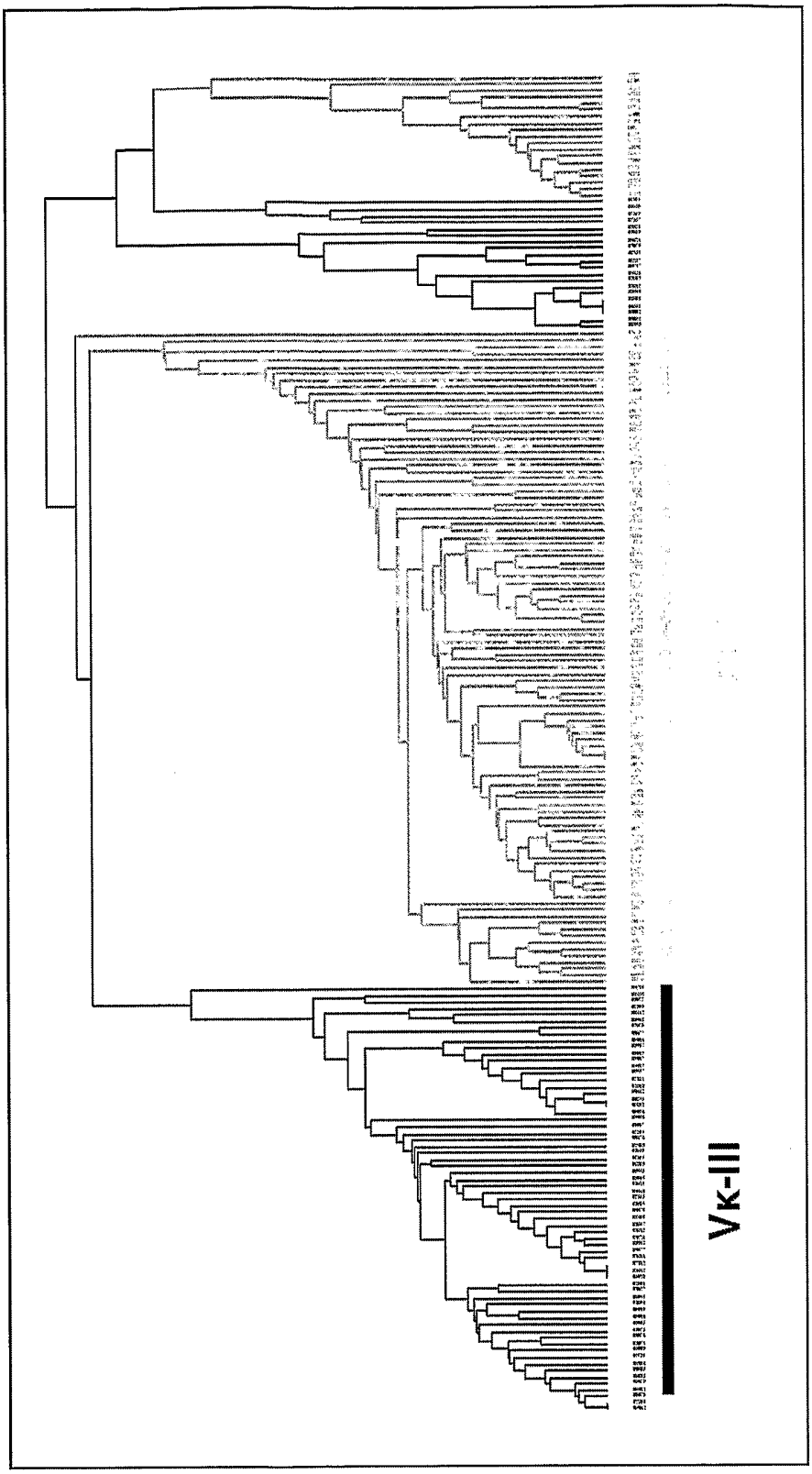


Fig. 38

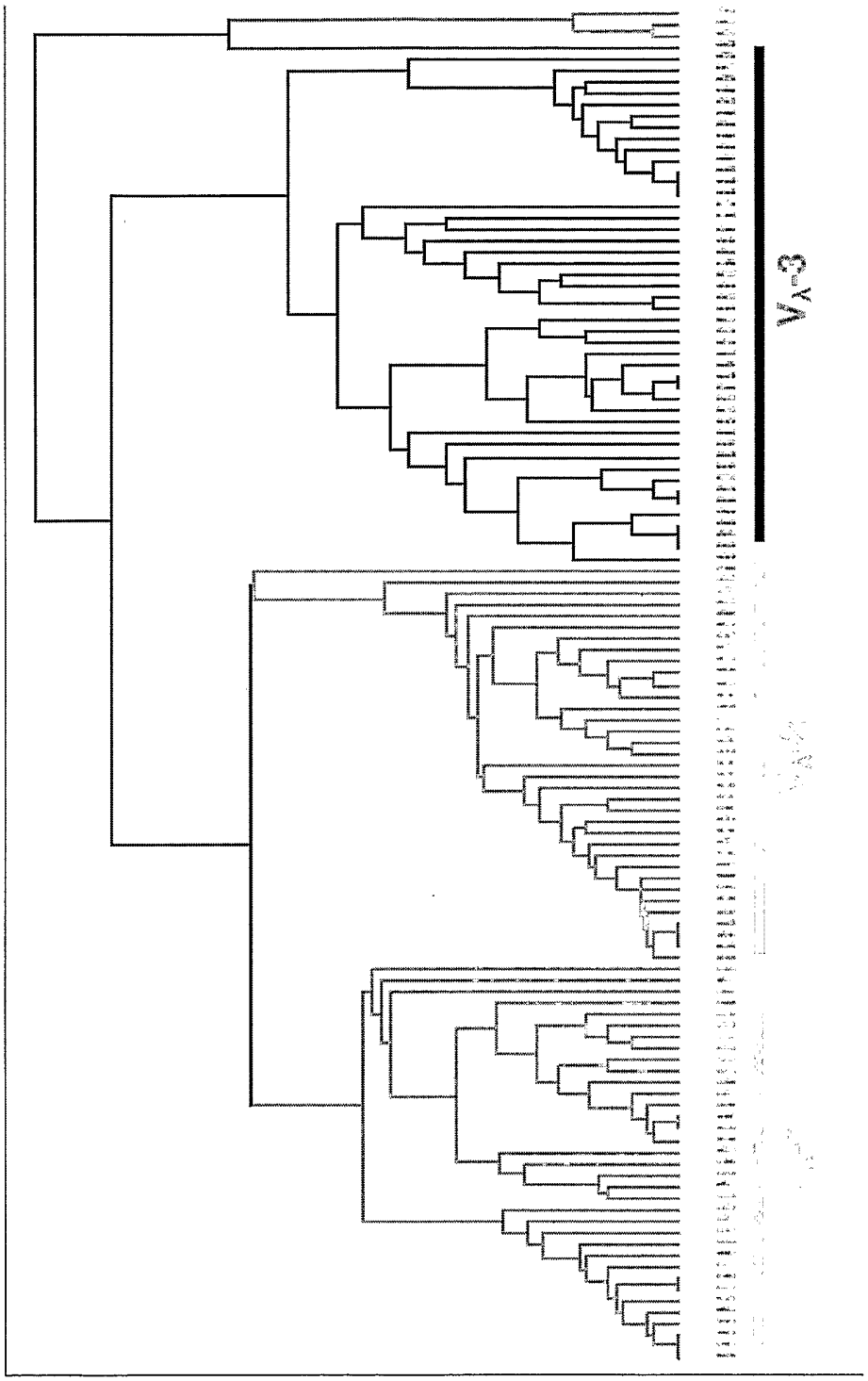


Fig. 39

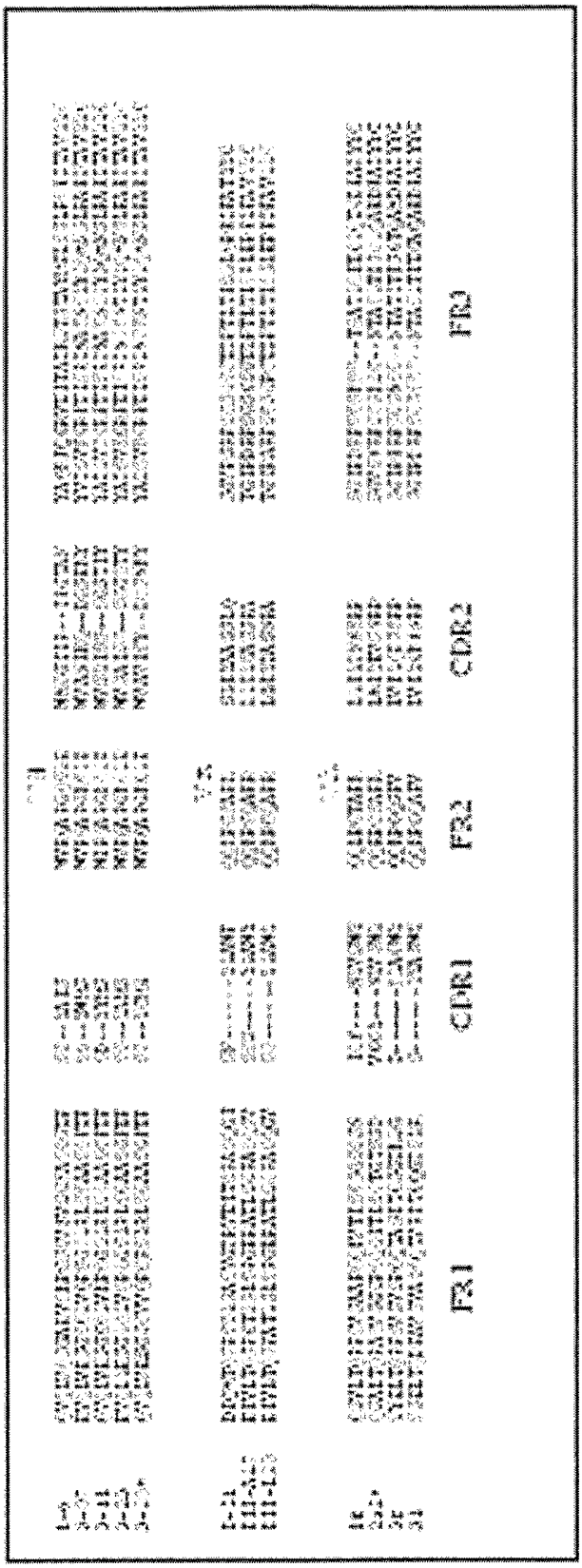


Fig. 40

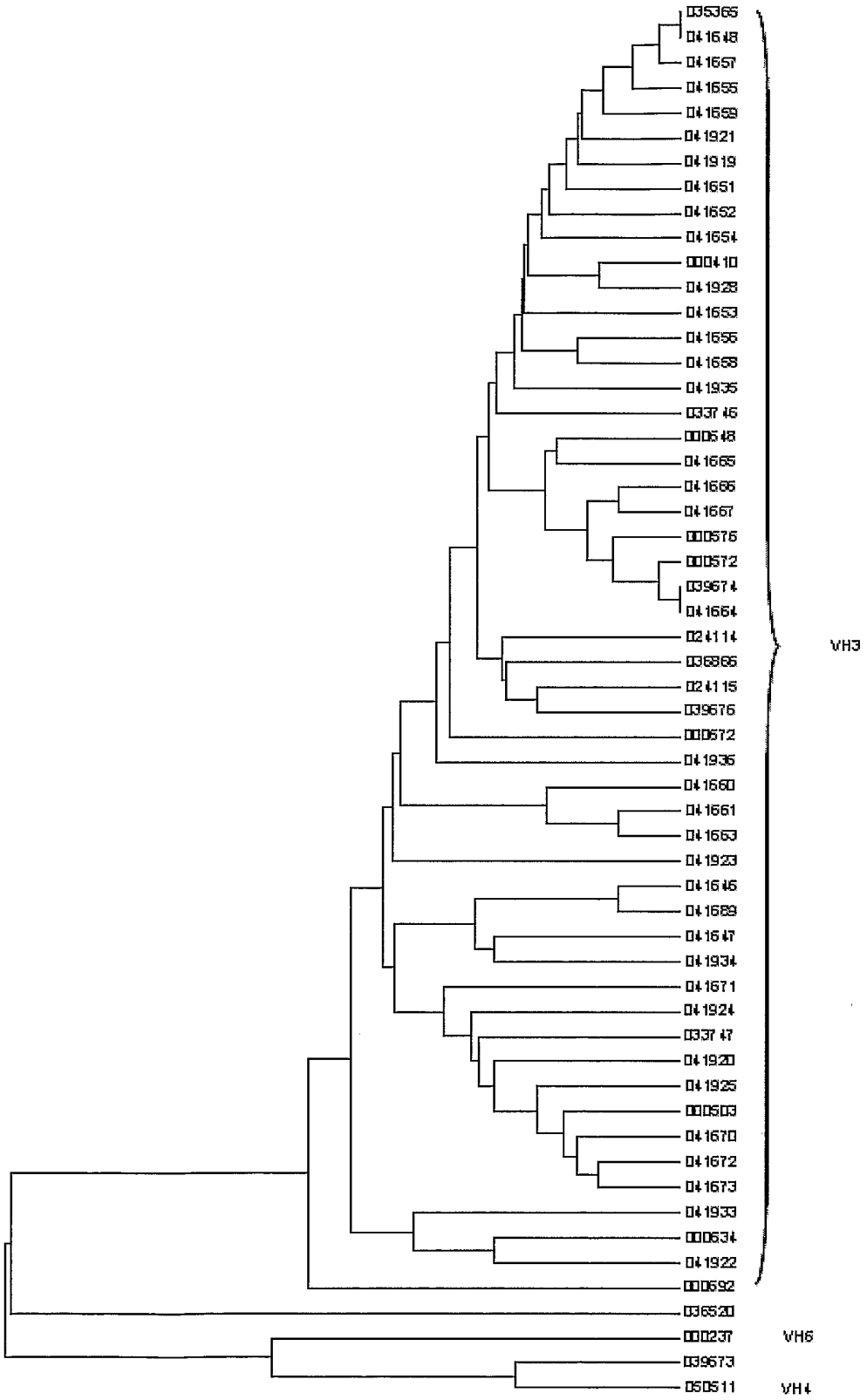

035365
041648
041657
041655
041659
041921
041919
041651
041652
041654
000410
041928
041653
041656
041658
041935
033745
000648
041665
041666
041667
000576
000572
039674
041664
024114
036866
024115
039676
000672
041936
041660
041661
041663
041923
041646
041669
041647
041934
041671
041924
033747
041920
041925
000503
041670
041672
041673
041933
000634
041922
000692
036520
000237
039673
050511
VH3
VH6
VH4

Fig. 41

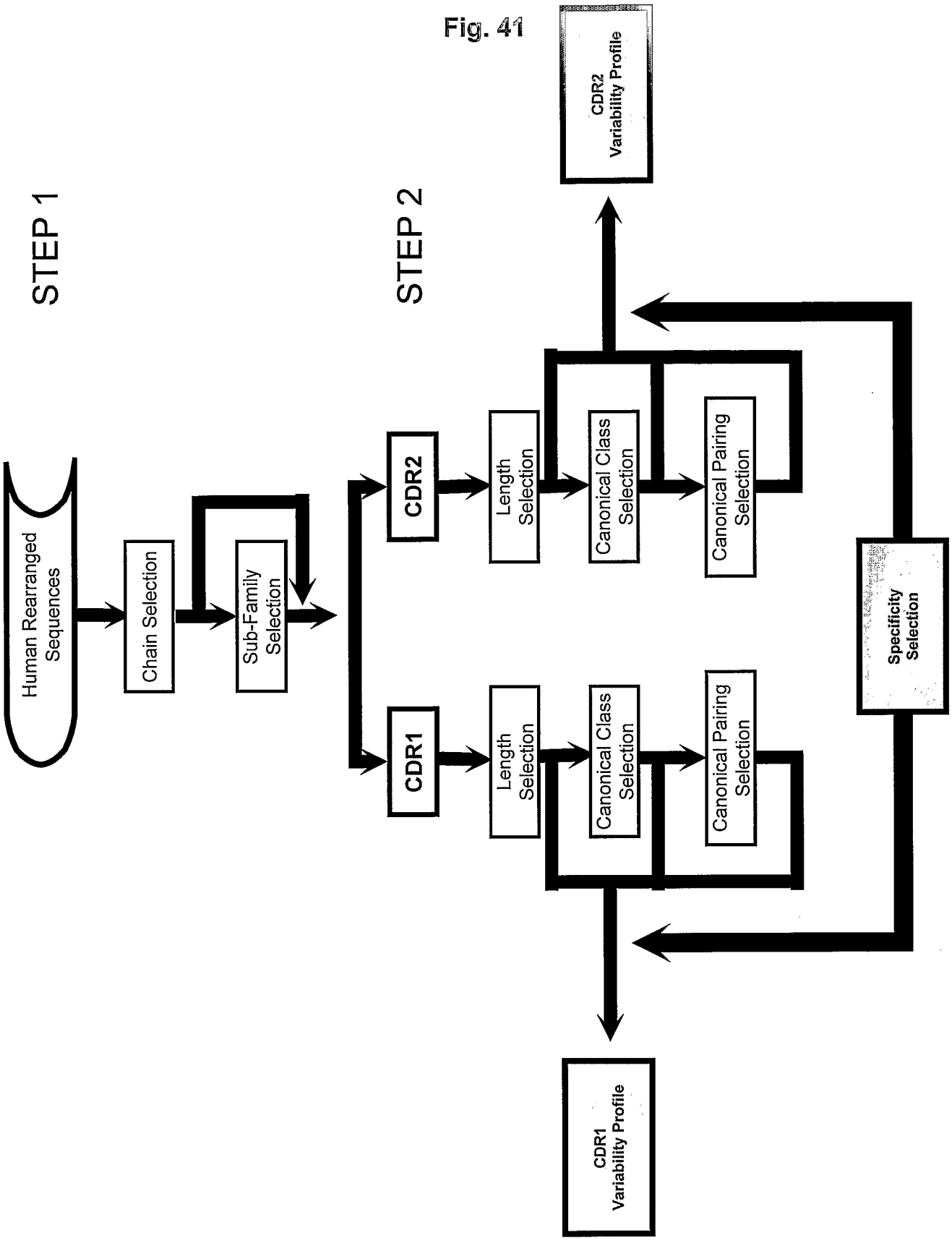

STEP 1
STEP 2
Human Rearranged Sequences
Chain Selection
Sub-Family Selection
CDR2
CDR1
Length Selection
Canonical Class Selection
Canonical Pairing Selection
Length Selection
Canonical Class Selection
Canonical Pairing Selection
Specificity Selection
CDR2 Variability Profile
CDR1 Variability Profile

Fig. 42

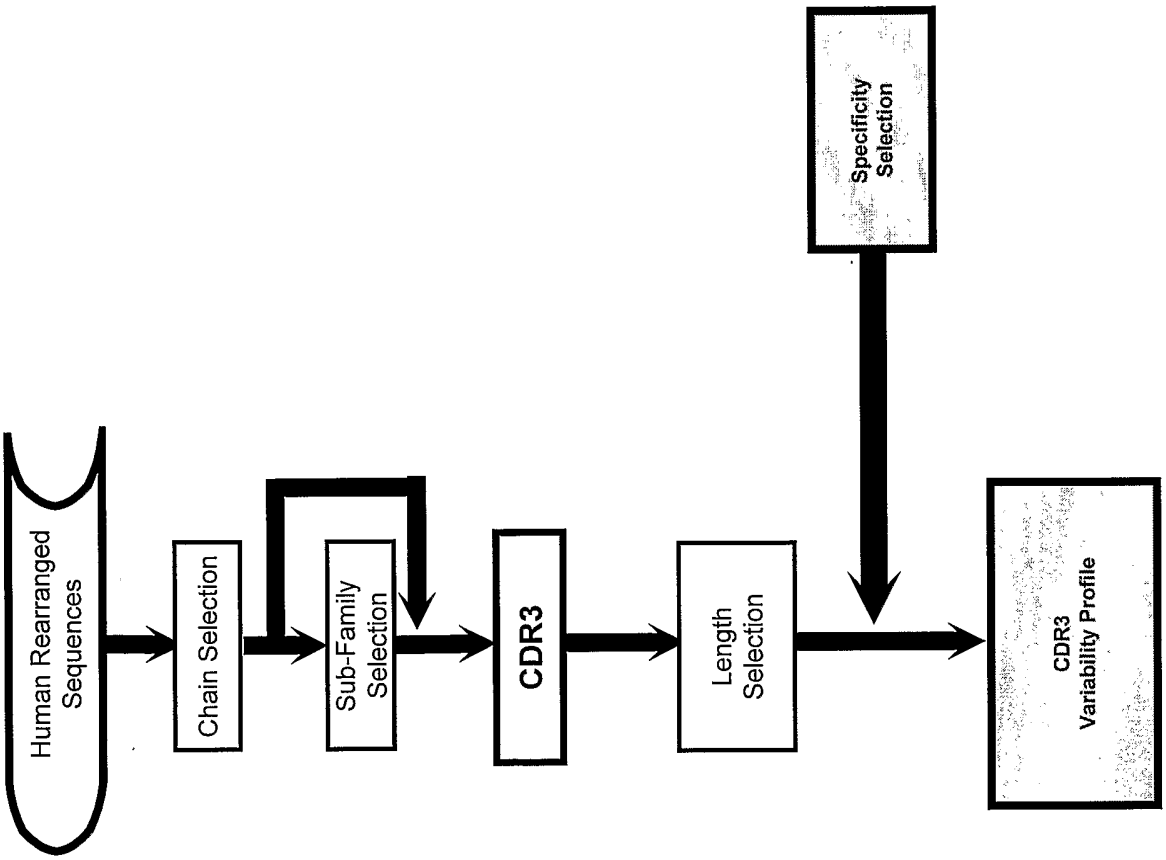

Human Rearranged Sequences
Chain Selection
Sub-Family Selection
CDR3
Length Selection
Specificity Selection
CDR3 Variability Profile

Fig. 43

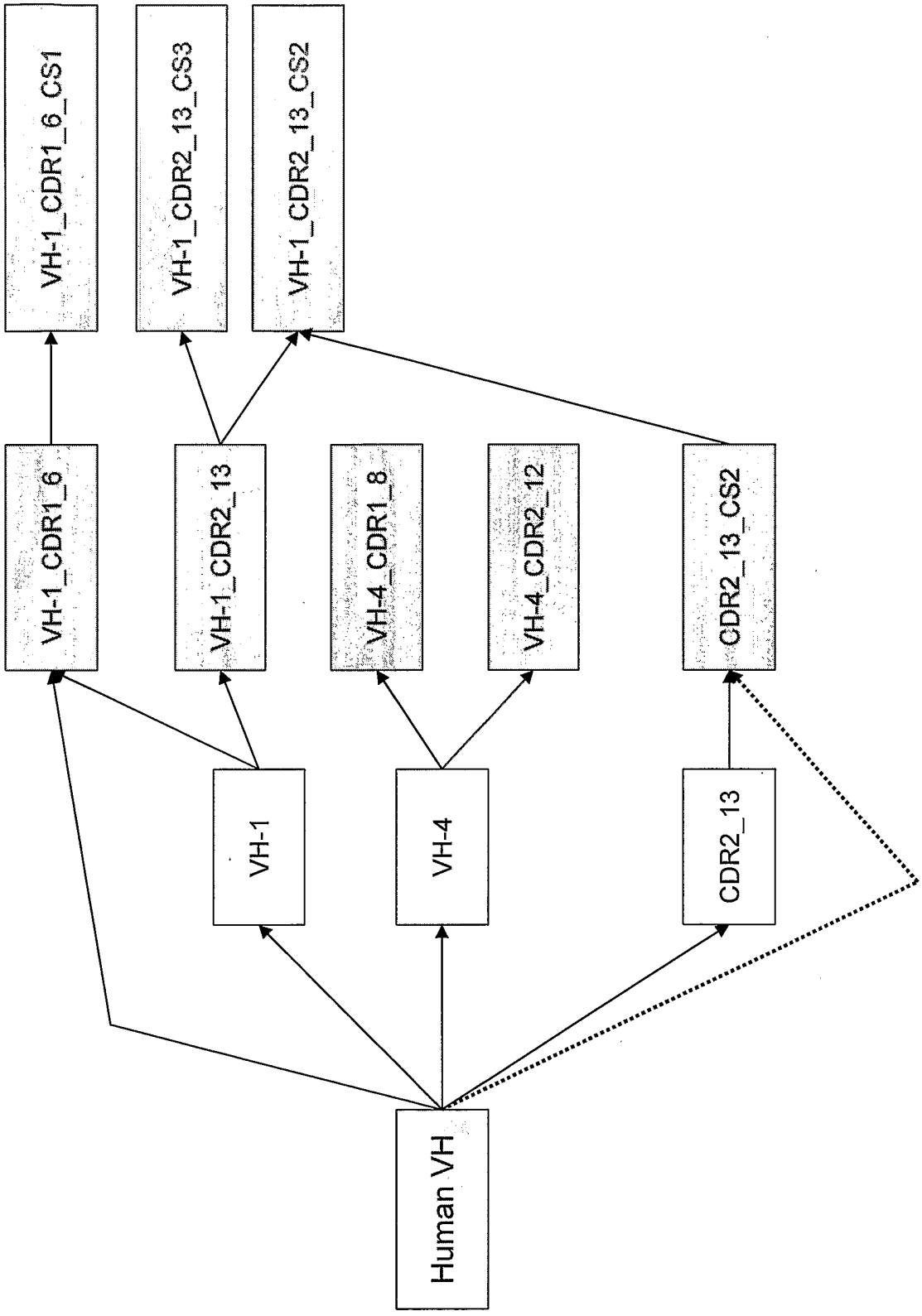

VH-1_CDR1_6_CS1
VH-1_CDR2_13_CS3
VH-1_CDR2_13_CS2
VH-1_CDR1_6
VH-1_CDR2_13
VH-4_CDR1_8
VH-4_CDR2_12
CDR2_13_CS2
VH-1
VH-4
CDR2_13
Human VH

Fig. 44

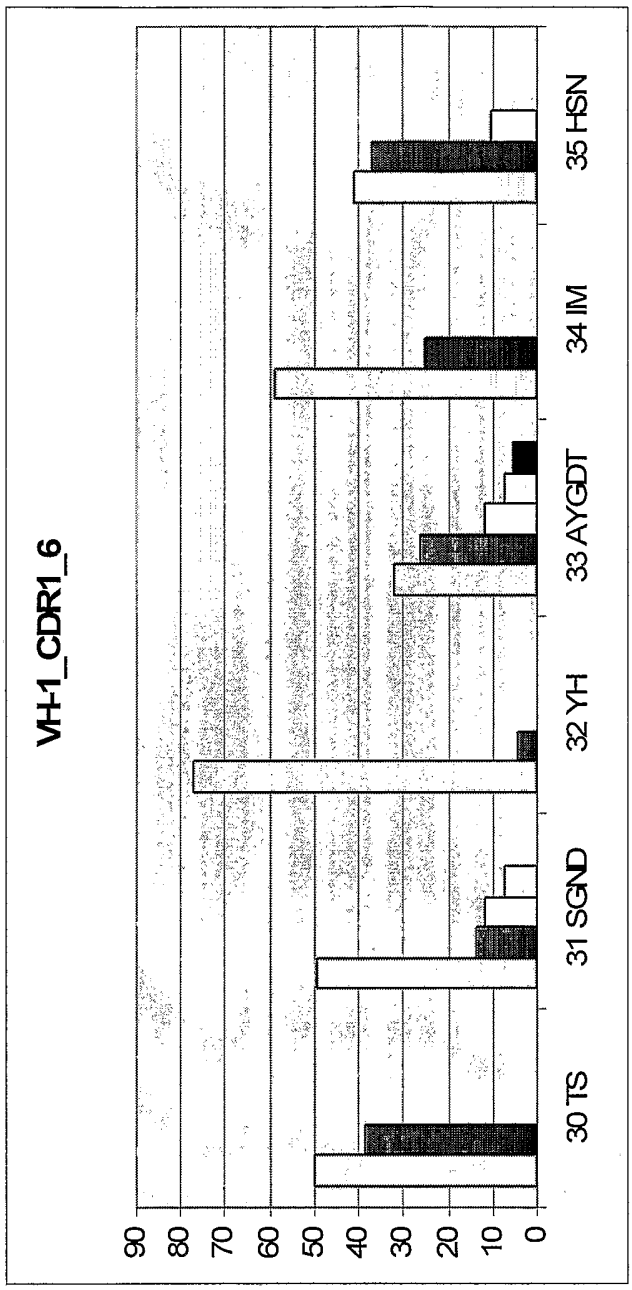

VH-1_CDR1_6
90
80
70
60
50
40
30
20
10
0
30 TS
31 SGND
32 YH
33 AYGDT
34 IM
35 HSN

Fig. 45

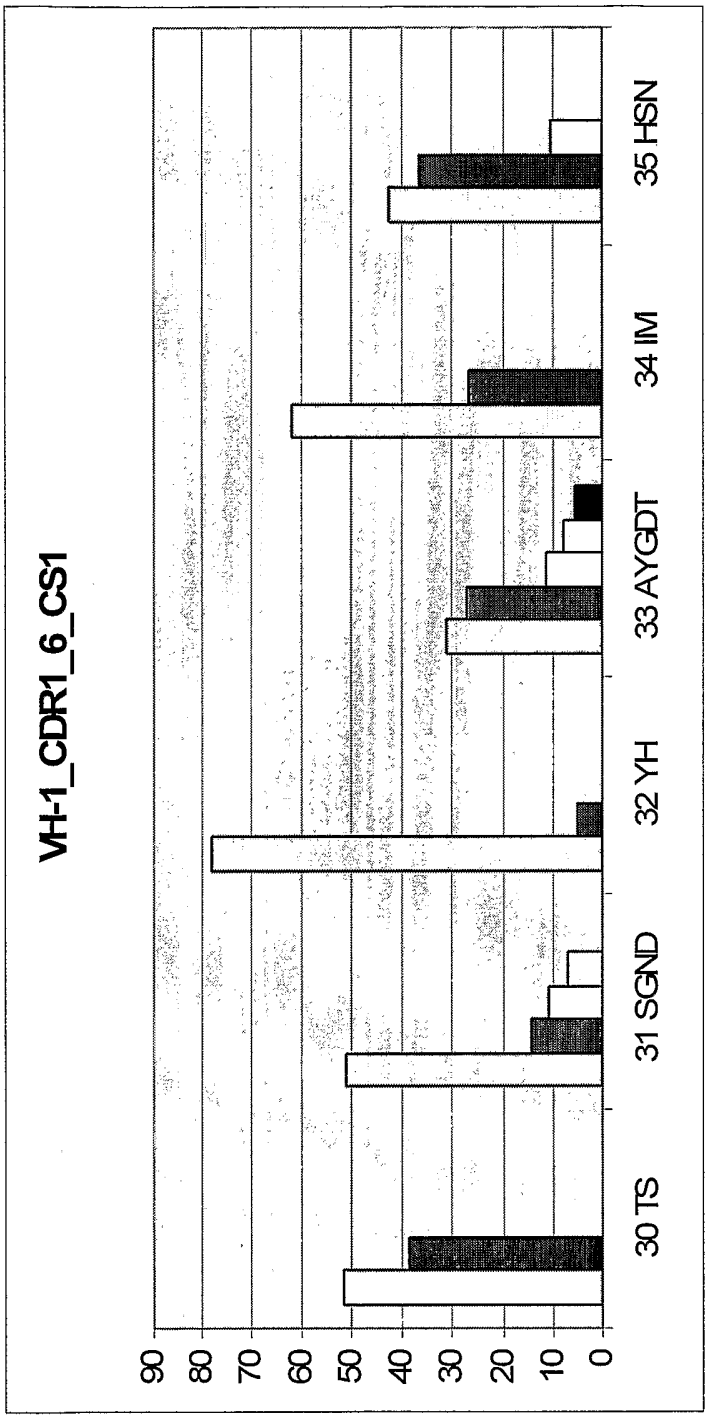

VH-1_CDR1_6_CS1
90
80
70
60
50
40
30
20
10
0
30 TS
31 SGND
32 YH
33 AYGDT
34 IM
35 HSN

Fig. 46

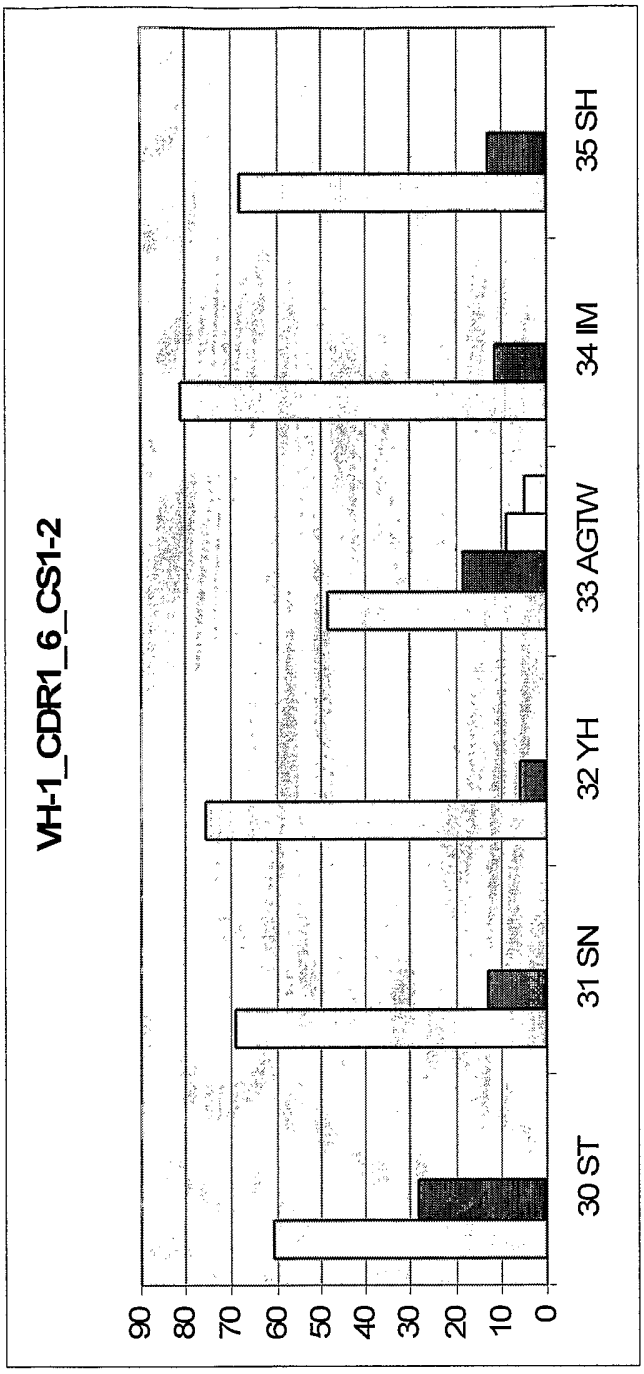

VH-1_CDR1_6_CS1-2
90
80
70
60
50
40
30
20
10
0
30 ST
31 SN
32 YH
33 AGTW
34 IM
35 SH

Fig. 47

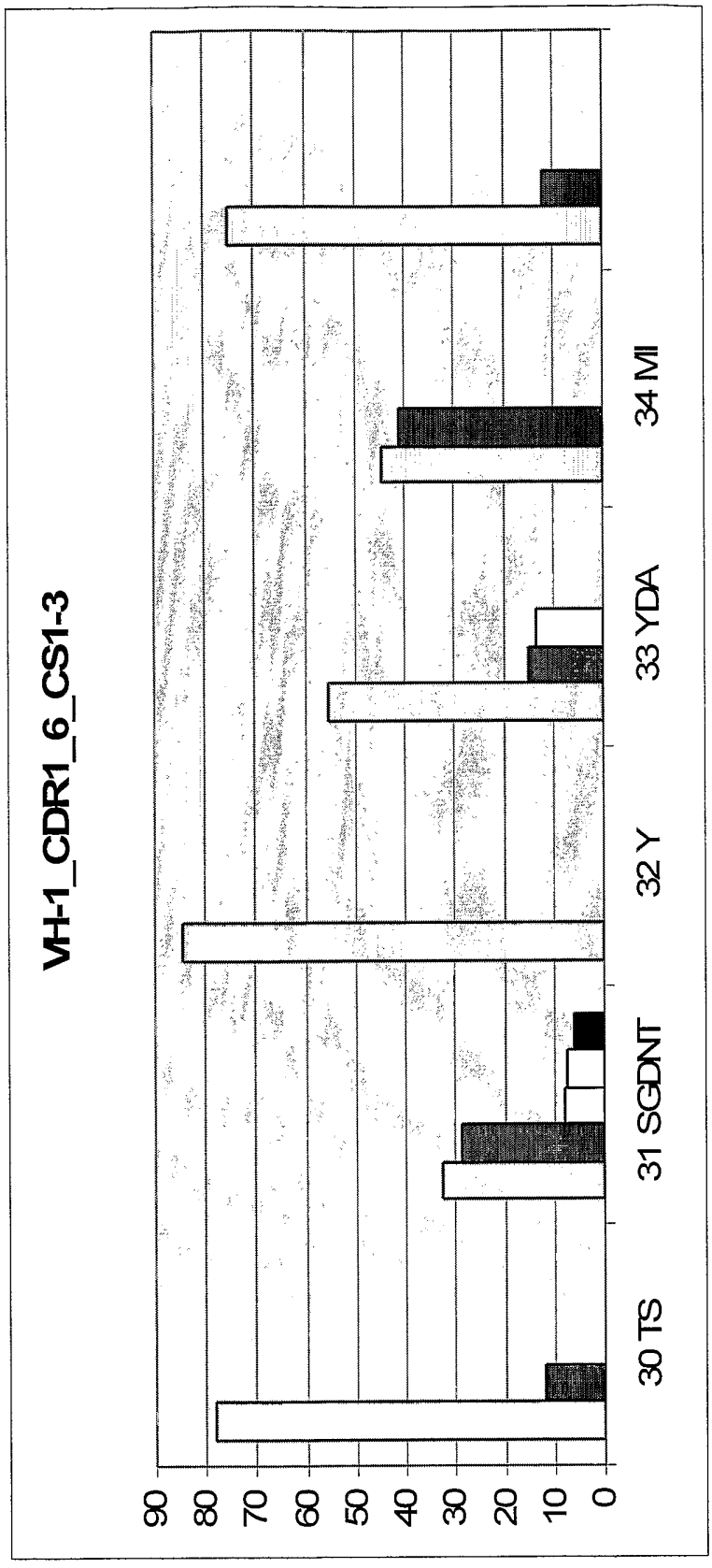

VH-1_CDR1_6_CS1-3
90
80
70
60
50
40
30
20
10
0
30 TS
31 SGDNT
32 Y
33 YDA
34 MI

Fig. 48

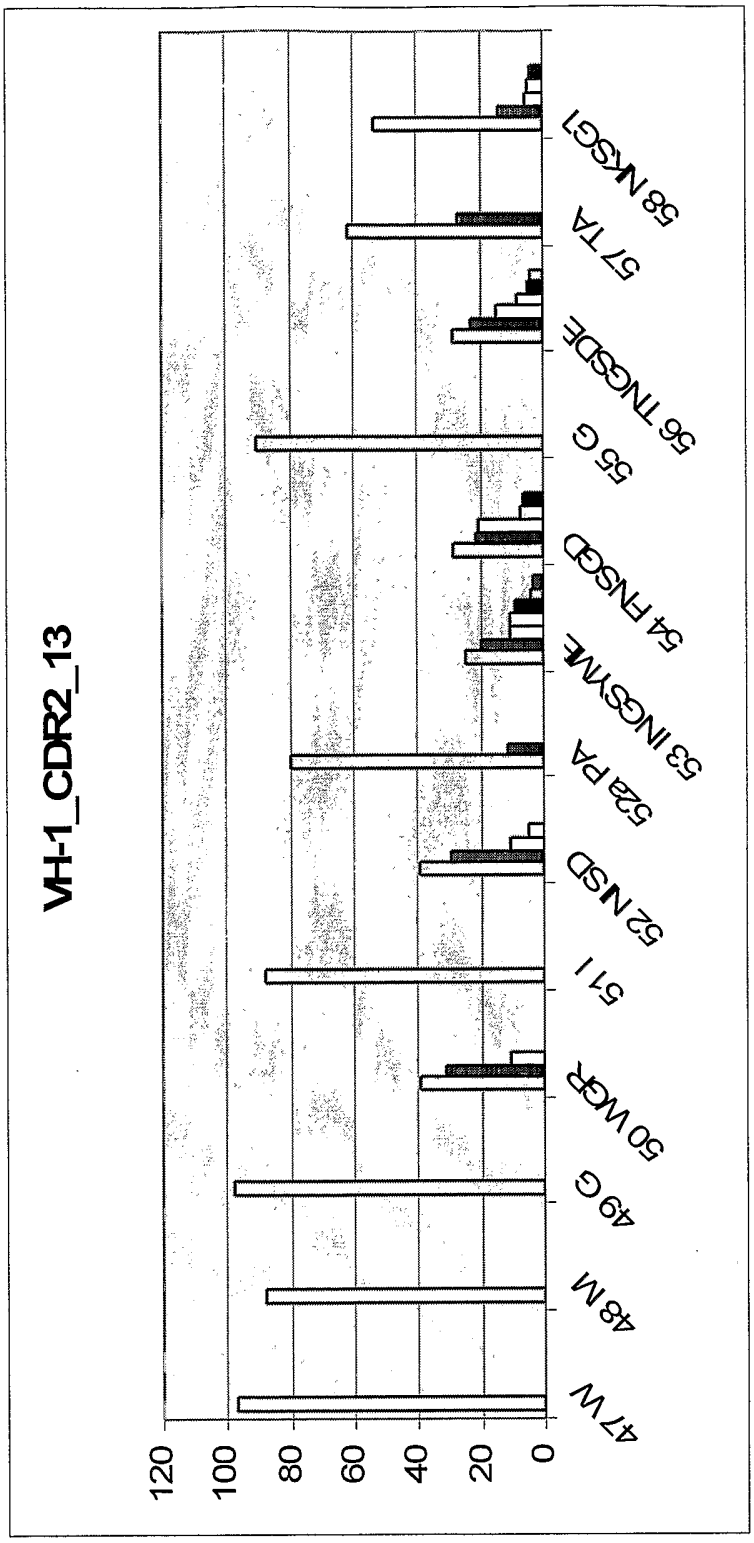

VH-1_CDR2_13
120
100
80
60
40
20
0
47W
48M
49G
50WGR
51I
52NSD
52aPA
53INGSVME
54FNSGD
55G
56TNGSDE
57TA
58NKSGT

Fig. 49

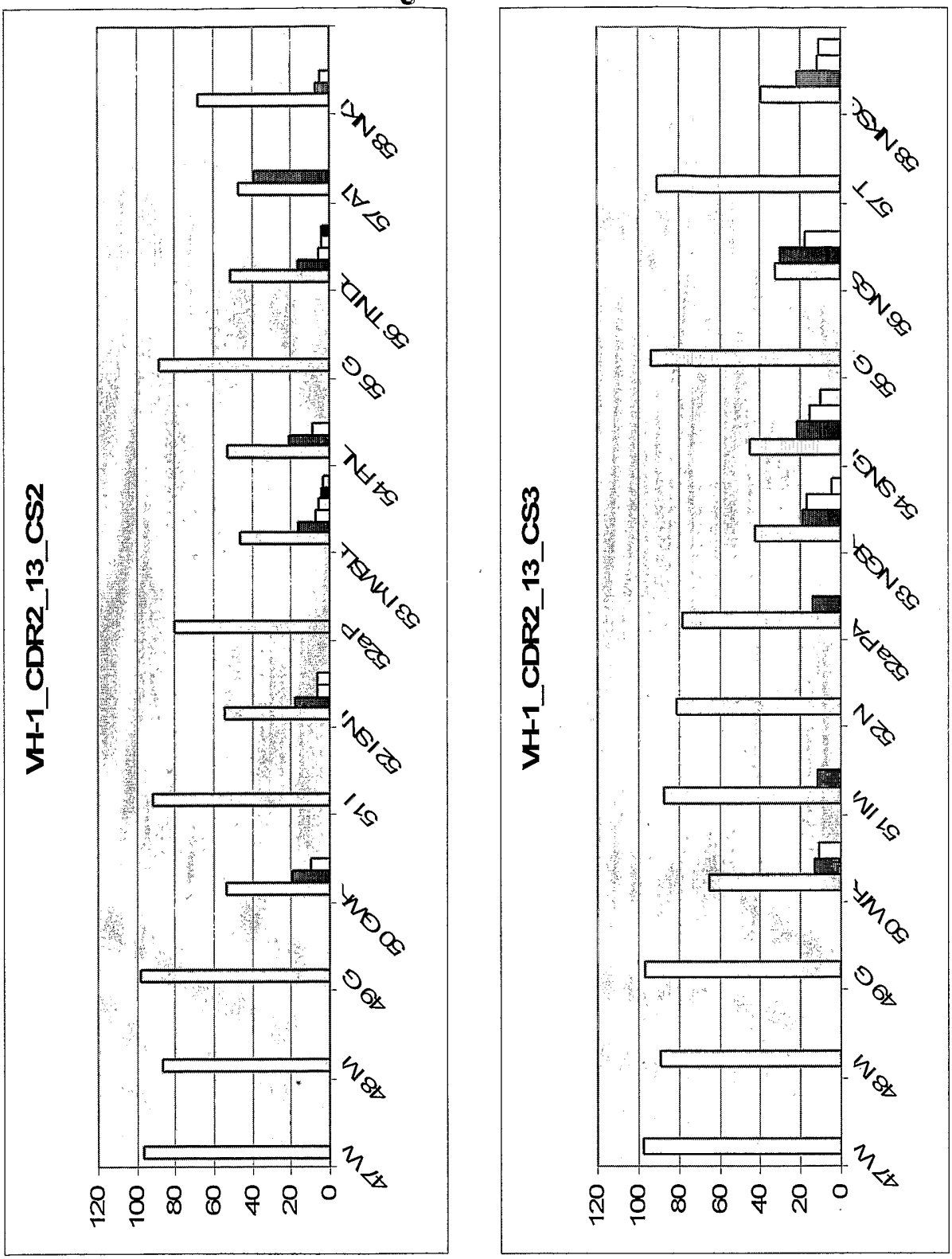

VH-1_CDR2_13_CS2
120
100
80
60
40
20
0
47W
48M
49G
50QMK
51I
52ISN
52aP
53YMEL
54FNL
55G
56TNDK
57AT
58NKT
VH-1_CDR2_13_CS3
120
100
80
60
40
20
0
47W
48M
49G
50WFK
51IM
52N
52aPA
53NQSK
54SNGI
55G
56NQS
57T
58NKS

Fig. 50

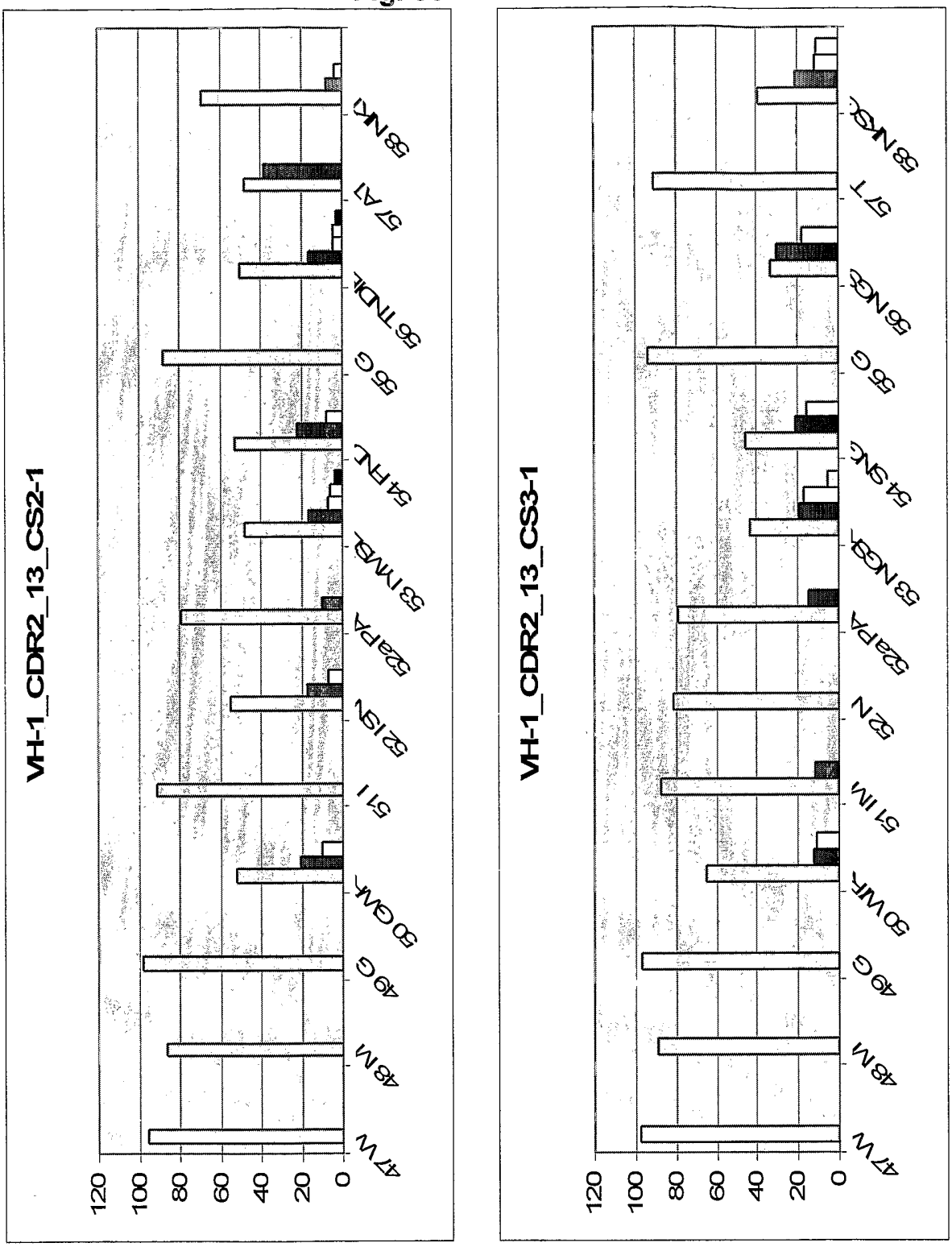
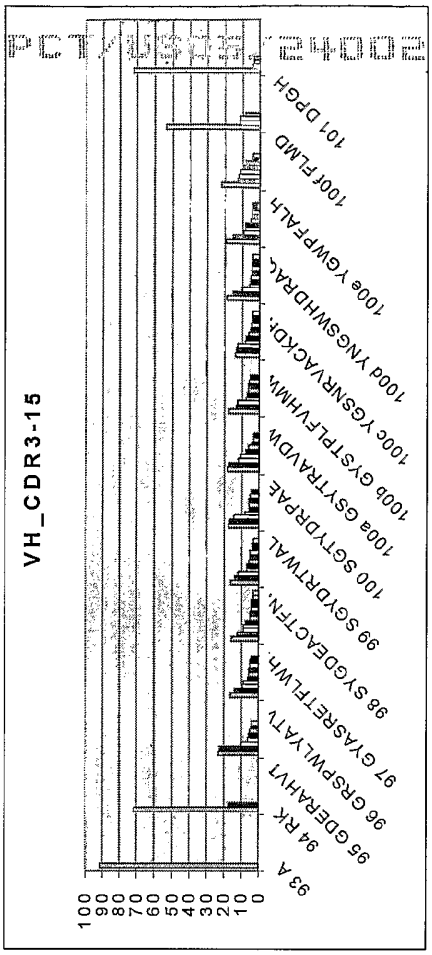
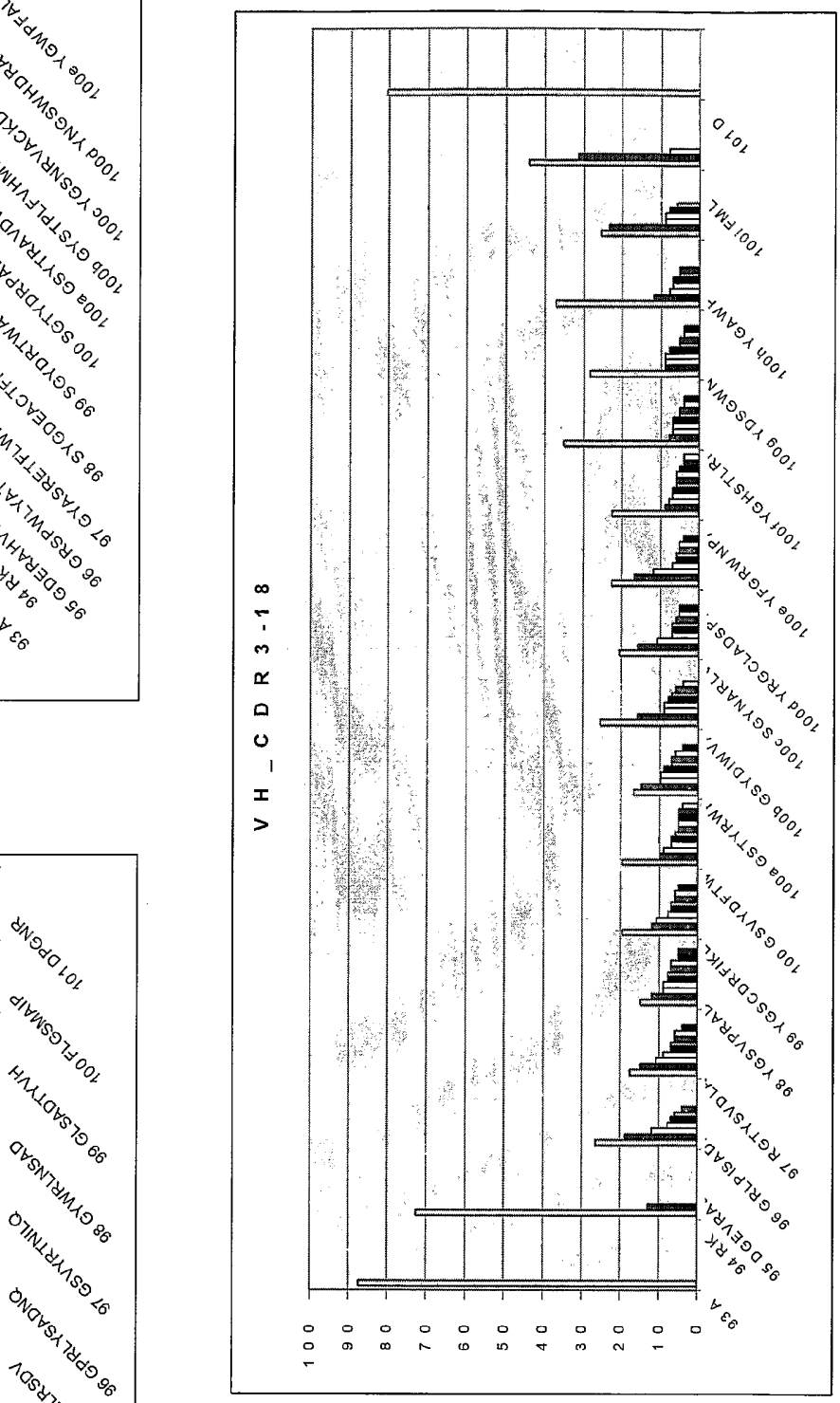
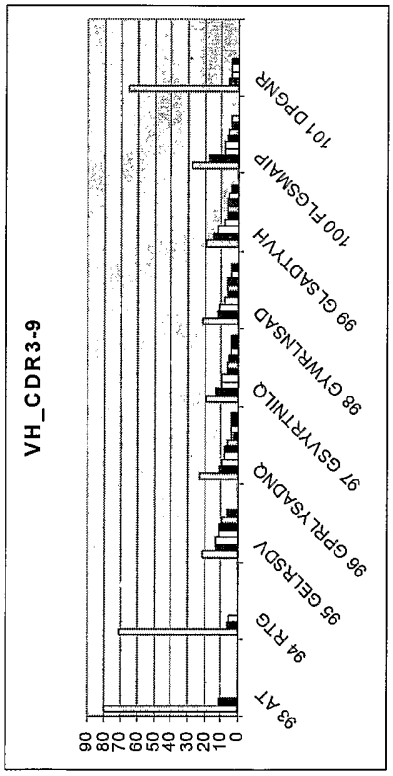

VH-1_CDR2_13_CS2-1
120
100
80
60
40
20
0
47W
48M
49G
50QK
51I
52IGV
52aPA
53YNS
54FKL
55G
56TNDK
57AI
58NKQ
VH-1_CDR2_13_CS3-1
120
100
80
60
40
20
0
47W
48M
49G
50WK
51IM
52N
52aPA
53NQSK
54SNQ
55G
56NQS
57T
58NQS

Fig. 51


VH_CDR3-15
100
90
80
70
60
50
40
30
20
10
0
93 A
94 RK
95 GDERAHVT
96 GRSPWLYATV
97 GYASRETFLWH
98 SYGDEACTFN
99 SGYDRTWAL
100 SGTYDRPAE
100a GSYTRAVDV
100b GYSTPLFVHM
100c YGSNRVACKD
100d YNGSWHDRAL
100e YGWPFALH
100f FLMD
101 DPGH


VH_CDR3-18
100
90
80
70
60
50
40
30
20
10
0
93 A
94 RK
95 DGEVRA
96 GRLPISAD
97 RGTYSVDLA
98 YGSVPRAL
99 YGSCDRFIKL
100 GSVYDFTW
100a GSTYRWI
100b GSYDIWV
100c SGYNARL
100d YRGCLADSP
100e YFGRWNP
100f YGHSTLR
100g YDSGWN
100h YGAWP
100i FML
101 D


VH_CDR3-9
90
80
70
60
50
40
30
20
10
0
93 AT
94 RTG
95 GELRSDV
96 GPRLYSADNQ
97 GSVYRTNILQ
98 GYWRLNSAD
99 GLSADTYVH
100 FLGSMAIP
101 DPGNR