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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 March 2009 (19.03.2009)

PCT

(10) International Publication Number
WO 2009/036379 A2

(51) International Patent Classification:
C07K 16/00 (2006.01) *C12N 15/13* (2006.01)

(21) International Application Number:
PCT/US2008/076300

(22) International Filing Date:
12 September 2008 (12.09.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/993,785 14 September 2007 (14.09.2007) US

(71) Applicant *(for all designated States except US)*: **ADIMAB, INC.** [US/US]; 16 Cavendish Court, Lebanon, NH 03766 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **VASQUEZ, Maximiliano** [US/US]; 3813 Louis Road, Palo Alto, CA 94303 (US). **FELDHAUS, Michael** [US/US]; 4 Knob Hill, Grantham, NH 03753 (US). **GERNGROSS, Tillman, U.** [AT/US]; 2 Parkway, Hanover, NH 03755 (US). **WITTRUP, K., Dane** [US/US]; 48 Woodlawn Drive, Chestnut Hill, MA 02467 (US).

(74) Agents: SMITH, DeAnn, F. et al.; Foley Hoag, LLP, Patent Group, 155 Seaport Boulevard, Boston, MA 02210-2600 (US).

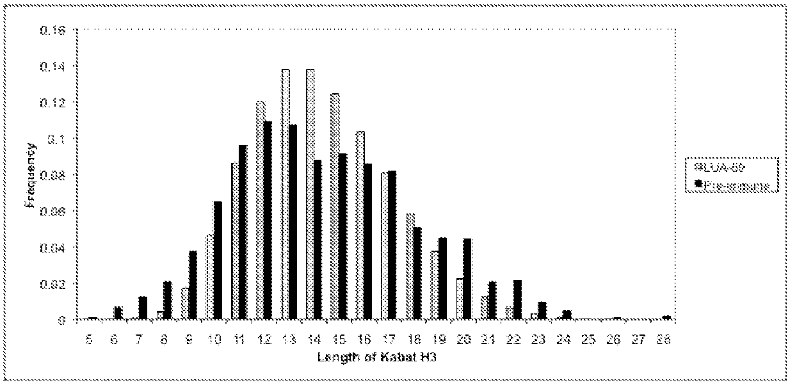
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)***:** AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG,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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)***:** ARIPO (BW, GH, GM, KE, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, 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*

(54) Title: RATIONALLY DESIGNED, SYNTHETIC ANTIBODY LIBRARIES AND USES THEREFOR

Figure 23



(57) Abstract: The present invention overcomes the inadequacies inherent in the known methods for generating libraries of antibody-encoding polynucleotides by specifically designing the libraries with directed sequence and length diversity. The libraries are designed to reflect the preimmune repertoire naturally created by the human immune system and are based on rational design informed by examination of publicly available databases of human antibody sequences.

RATIONALLY DESIGNED, SYNTHETIC ANTIBODY LIBRARIES AND USES THEREFOR

RELATED APPLICATION

This application claims priority to U.S. provisional application serial number 60/993,785, filed on September 14, 2007, incorporated herein in its entirety by this reference.

BACKGROUND OF THE INVENTION

Antibodies have profound relevance as research tools and in diagnostic and therapeutic applications. However, the identification of useful antibodies is difficult and once identified, antibodies often require considerable redesign or ‘humanization’ before they are suitable for therapeutic applications.

Previous methods for identifying desirable antibodies have typically involved phage display of representative antibodies, for example human libraries derived by amplification of nucleic acids from B cells or tissues, or, alternatively, synthetic libraries. However, these approaches have limitations. For example, most human libraries known in the art contain only the antibody sequence diversity that can be experimentally captured or cloned from the source (*e.g.,* B cells). Accordingly, the human library may completely lack or under-represent certain useful antibody sequences. Synthetic or consensus libraries known in the art have other limitations, such as the potential to encode non-naturally occurring (*e.g.,* non-human) sequences that have the potential to be immunogenic. Moreover, certain synthetic libraries of the art suffer from at least one of two limitations: (1) the number of members that the library can theoretically contain (*i.e.,* theoretical diversity) may be greater than the number of members that can actually be synthesized, and (2) the number of members actually synthesized may be so great as to preclude screening of each member in a physical realization of the library, thereby decreasing the probability that a library member with a particular property may be isolated.

For example, a physical realization of a library (*e.g.,* yeast display, phage display, ribosomal display, etc.) capable of screening 10^{12} library members will only sample about 10% of the sequences contained in a library with 10^{13} members. Given a median CDRH3 length of about 12.7 amino acids (Rock *et al.*, J. Exp. Med., 1994, 179:323-328), the number of theoretical sequence variants in CDRH3 alone is about $20^{12.7}$, or about 3.3×10^{16} variants. This number does not account for known variation that occurs in CDRH1 and CDRH2, heavy chain framework regions, and pairing with different light chains, each of which also exhibit variation in their respective CDRL1, CDRL2, and CDRL3. Finally, the antibodies isolated from these libraries are often not amenable to rational affinity maturation techniques to improve the binding of the candidate molecule.

Accordingly, a need exists for smaller (*i.e.,* able to be synthesized and physically realizable) antibody libraries with directed diversity that systematically represent candidate antibodies that are non-immunogenic (*i.e.,* more human) and have desired properties (*e.g.,* the ability to recognize a broad variety of antigens). However, obtaining such libraries requires balancing the competing objectives of restricting the sequence diversity represented in the library (to enable synthesis and physical realization, potentially with oversampling, while limiting the introduction of non-human sequences) while maintaining a level of diversity sufficient to recognize a broad variety of antigens. Prior to the instant invention, it was known in the art that "[al]though libraries containing heavy chain CDR3 length diversity have been reported, it is impossible to synthesize DNA encoding both the sequence and the length diversity found in natural heavy chain CDR3 repertoires" (Hoet *et al.,* Nat. Biotechnol., 2005, 23: 344, incorporated by reference in its entirety).

Therefore, it would be desirable to have antibody libraries which (a) can be readily synthesized, (b) can be physically realized and, in certain cases, oversampled, (c) contain sufficient diversity to recognize all antigens recognized by the preimmune human repertoire (*i.e.,* before negative selection), (d) are non-immunogenic in humans (*i.e.,* comprise sequences of human origin), and (e) contain CDR length and sequence diversity, and framework diversity, representative of naturally-occurring human antibodies. Embodiments of the instant invention at least provide, for the first time, antibody libraries that have these desirable features.

SUMMARY OF THE INVENTION

The present invention is relates to, at least, synthetic polynucleotide libraries, methods of producing and using the libraries of the invention, kits and computer readable forms including the libraries of the invention. In some embodiments, the libraries of the invention are designed to reflect the preimmune repertoire naturally created by the human immune system and are based on rational design informed by examination of publicly available databases of human antibody sequences. It will be appreciated that certain non-limiting embodiments of the invention are described below. As described throughout the specification, the invention encompasses many other embodiments as well.

In certain embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode at least 10^6 unique antibody CDRH3 amino acid sequences comprising:

(i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

(ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N2 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

(iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

In other embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode at least about 10^6 unique antibody CDRH3 amino acid sequences comprising:

(i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein:

(a) the most N-terminal N1 amino acid, if present, is selected from a group consisting of R, G, P, L, S, A, V, K, I, Q, T and D;

(b) the second most N-terminal N1 amino acid, if present, is selected from a group consisting of G, P, R, S, L, V, E, A, D, I, T and K; and

(c) the third most N-terminal N1 amino acid, if present, is selected from the group consisting of G, R, P, S, L, A, V, T, E, D, K and F;

(ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein:

(a) the most N-terminal N2 amino acid, if present, is selected from a group consisting of G, P, R, L, S, A, T, V, E, D, F and H;

(b) the second most N-terminal N2 amino acid, if present, is selected from a group consisting of G, P, R, S, T, L, A, V, E, Y, D and K; and

(c) the third most N-terminal N2 amino acid, if present, is selected from the group consisting of G, P, S, R, L, A, T, V, D, E, W and Q; and

(iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

In still other embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode at least about 10^6 unique antibody CDRH3 amino acid sequences that are at least about 80% identical to an amino acid sequence represented by the following formula:

[X]-[N1]-[DH]-[N2]-[H3-JH], wherein:

(i) X is any amino acid residue or no amino acid residue;

(ii) N1 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof;

(iii) DH is an amino acid sequence selected from the group consisting of all possible reading frames that do not include a stop codon encoded by IGHD1-1, IGHD1-20, IGHD1-26, IGHD1-7, IGHD2-15, IGHD2-2, IGHD2-21, IGHD2-8, IGHD3-10, IGHD3-16, IGHD3-22, IGHD3-3, IGHD3-9, IGHD4-17, IGHD4-23, IGHD4-4, IGHD-4-11, IGHD5-12, IGHD5-24, IGHD5-5, IGHD-5-18, IGHD6-13, IGHD6-19, IGHD6-25, IGHD6-6, and IGHD7-27, and N- and C-terminal truncations thereof;

(iv) N2 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof; and

(v) H3-JH is an amino acid sequence selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV, or a sequence of at least 80% identity to any of them.

In still another embodiment, the invention comprises wherein said library consists essentially of a plurality of polynucleotides encoding CDRH3 amino acid sequences that are at least about 80% identical to an amino acid sequence represented by the following formula:

[X]-[N1]-[DH]-[N2]-[H3-JH], wherein:

(i) X is any amino acid residue or no amino acid residue;

(ii) N1 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof;

(iii) DH is an amino acid sequence selected from the group consisting of all possible reading frames that do not include a stop codon encoded by IGHD1-1, IGHD1-20, IGHD1-26, IGHD1-7, IGHD2-15, IGHD2-2, IGHD2-21, IGHD2-8, IGHD3-10, IGHD3-16, IGHD3-22, IGHD3-3, IGHD3-9, IGHD4-17, IGHD4-23, IGHD4-4, IGHD-4-11, IGHD5-12, IGHD5-24, IGHD5-5, IGHD-5-18, IGHD6-13, IGHD6-19, IGHD6-25, IGHD6-6, and IGHD7-27, and N- and C-terminal truncations thereof;

(iv) N2 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof; and

(v) H3-JH is an amino acid sequence selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV, or a sequence of at least 80% identity to any of them.

In another embodiment, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode one or more full length antibody heavy chain sequences, and wherein the CDRH3 amino acid sequences of the heavy chain comprise:

(i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

(ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N2 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

(iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

The following embodiments may be applied throughout the embodiments of the instant invention. In one aspect, one or more CDRH3 amino acid sequences further comprise an N-terminal tail residue. In still another aspect, the N-terminal tail residue is selected from the group consisting of G, D, and E.

In yet another aspect, the N1 amino acid sequence is selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof. In certain other aspects, the N1 amino acid sequence may be of about 0 to about 5 amino acids.

In yet another aspect, the N2 amino acid sequence is selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE,

QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof. In certain other aspects, the N2 sequence may be of about 0 to about 5 amino acids.

In yet another aspect, the H3-JH amino acid sequence is selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV.

In other embodiments, the invention comprises a library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+1 pairs in the library is within the ranges specified below:

Tyr-Tyr in an amount from about 2.5% to about 6.5%;

Ser-Gly in an amount from about 2.5% to about 4.5%;

Ser-Ser in an amount from about 2% to about 4%;

Gly-Ser in an amount from about 1.5% to about 4%;

Tyr-Ser in an amount from about 0.75% to about 2%;

Tyr-Gly in an amount from about 0.75% to about 2%; and

Ser-Tyr in an amount from about 0.75% to about 2%.

In still other embodiments, the invention comprises a library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+2 pairs in the library is within the ranges specified below:

Tyr-Tyr in an amount from about 2.5% to about 4.5%;

Gly-Tyr in an amount from about 2.5% to about 5.5%;

Ser-Tyr in an amount from about 2% to about 4%;

Tyr-Ser in an amount from about 1.75% to about 3.75%;

Ser-Gly in an amount from about 2% to about 3.5%;

Ser-Ser in an amount from about 1.5% to about 3%;

Gly-Ser in an amount from about 1.5% to about 3%; and

Tyr-Gly in an amount from about 1% to about 2%.

In another embodiment, the invention comprises a library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+3 pairs in the library is within the ranges specified below:

Gly-Tyr in an amount from about 2.5% to about 6.5%;

Ser-Tyr in an amount from about 1% to about 5%;

Tyr-Ser in an amount from about 2% to about 4%;

Ser-Ser in an amount from about 1% to about 3%;

Gly-Ser in an amount from about 2% to about 5%; and

Tyr-Tyr in an amount from about 0.75% to about 2%.

In one aspect of the invention, at least 2, 3, 4, 5, 6, or 7 of the specified i – i+1 pairs in the library are within the specified ranges. In another aspect, the CDRH3 amino acid sequences are human. In yet another aspect, the polynucleotides encode at least about 10^6 unique CDRH3 amino acid sequences.

In other aspects of the invention, the polynucleotides further encode one or more heavy chain chassis amino acid sequences that are N-terminal to the CDRH3 amino acid sequences, and the one or more heavy chain chassis sequences are selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 94 encoded by IGHV1-2, IGHV1-3, IGHV1-8, IGHV1-18, IGHV1-24, IGHV1-45, IGHV1-46, IGHV1-58, IGHV1-69, IGHV2-5, IGHV2-26, IGHV2-70, IGHV3-7, IGHV3-9, IGHV3-11, IGHV3-13, IGHV3-15, IGHV3-20, IGHV3-21, IGHV3-23, IGHV3-30, IGHV3-33, IGHV3-43, IGHV3-48, IGHV3-49, IGHV3-53, IGHV3-64, IGHV3-66, IGHV3-72, IGHV3-73, IGHV3-74, IGHV4-4, IGHV4-28, IGHV4-31, IGHV4-34, IGHV4-39, IGHV4-59, IGHV4-61, IGHV4-B, IGHV5-51, IGHV6-1, and IGHV7-4-1, or a sequence of at least about 80% identity to any of them.

In another aspect, the polynucleotides further encode one or more FRM4 amino acid sequences that are C-terminal to the CDRH3 amino acid sequences, wherein the one or more FRM4 amino acid sequences are selected from the group consisting of a FRM4 amino acid sequence encoded by IGHJ1, IGHJ2, IGHJ3, IGHJ4, IGHJ5, and IGHJ6, or a sequence of at least about 80% identity to any of them. In still another aspect, the polynucleotides further encode one or more immunoglobulin heavy chain constant region amino acid sequences that are C-terminal to the FRM4 sequence.

In yet another aspect, the CDRH3 amino acid sequences are expressed as part of full-length heavy chains. In other aspects, the full-length heavy chains are selected from the group consisting of an IgG1, IgG2, IgG3, and IgG4, or combinations thereof. In one embodiment, the CDRH3 amino acid sequences are from about 2 to about 30, from about 8 to about 19, or from about 10 to about 18 amino acid residues in length. In other aspects, the synthetic polynucleotides of the library encode from about 10^6 to about 10^{14}, from about 10^7 to about 10^{13}, from about 10^8 to about 10^{12}, from about 10^9 to about 10^{12}, or from about 10^{10} to about 10^{12} unique CDRH3 amino acid sequences.

In certain embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody VKCDR3 amino acid sequences comprising about 1 to about 10 of the amino acids found at Kabat positions 89, 90, 91, 92, 93, 94, 95, 95A, 96, and 97, in selected VKCDR3 amino acid sequences derived from a particular IGKV or IGKJ germline sequence.

In one aspect, the synthetic polynucleotides encode one or more of the amino acid sequences listed in Table 33 or a sequence at least about 80% identical to any of them.

In some embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of unique antibody VKCDR3 amino acid sequences that are of at least about 80% identity to an amino acid sequence represented by the following formula:

[VK_Chassis]-[L3-VK]-[X]-[JK*], wherein:

(i) VK_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27,

IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30, IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-21, and IGKV6D-41, or a sequence of at least about 80% identity to any of them;

(ii) L3-VK is the portion of the VKCDR3 encoded by the IGKV gene segment; and

(iii) X is any amino acid residue; and

(iv) JK* is an amino acid sequence selected from the group consisting of sequences encoded by IGJK1, IGJK2, IGJK3, IGJK4, and IGJK5, wherein the first residue of each IGJK sequence is not present.

In still other aspects, X may be selected from the group consisting of F, L, I, R, W, Y, and P.

In certain embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of VλCDR3 amino acid sequences that are of at least about 80% identity to an amino acid sequence represented by the following formula:

[Vλ_Chassis]-[L3-Vλ]-[Jλ], wherein:

(i) Vλ_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGλV1-36, IGλV1-40, IGλV1-44, IGλV1-47, IGλV1-51, IGλV10-54, IGλV2-11, IGλV2-14, IGλV2-18, IGλV2-23, IGλV2-8, IGλV3-1, IGλV3-10, IGλV3-12, IGλV3-16, IGλV3-19, IGλV3-21, IGλV3-25, IGλV3-27, IGλV3-9, IGλV4-3, IGλV4-60, IGλV4-69, IGλV5-39, IGλV5-45, IGλV6-57, IGλV7-43, IGλV7-46, IGλV8-61, IGλV9-49, and IGλV10-54, or a sequence of at least about 80% identity to any of them;

(ii) L3-Vλ is the portion of the VλCDR3 encoded by the IGλV segment; and

(iii) Jλ is an amino acid sequence selected from the group consisting of sequences encoded by IGλJ1-01, IGλJ2-01, IGλJ3-01, IGλJ3-02, IGλJ6-01, IGλJ7-01, and IGλJ7-02, and wherein the first residue of each IGJλ sequence may or may not be deleted.

In further aspects, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody proteins comprising:

(i) a CDRH3 amino acid sequence of claim 1; and

(ii) a VKCDR3 amino acid sequence comprising about 1 to about 10 of the amino acids found at Kabat positions 89, 90, 91, 92, 93, 94, 95, 95A, 96, and 97, in selected VKCDR3 sequences derived from a particular IGKV or IGKJ germline sequence.

In still further aspects, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody proteins comprising:

(i) a CDRH3 amino acid sequence of claim 1; and

(ii) a VKCDR3 amino acid sequences of at least about 80% identity to an amino acid sequence represented by the following formula:

[VK_Chassis]-[L3-VK]-[X]-[JK*], wherein:

(a) VK_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27, IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30, IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-

21, and IGKV6D-41, or a sequence of at least about 80% identity to any of them;

(b) L3-VK is the portion of the VKCDR3 encoded by the IGKV gene segment; and

(c) X is any amino acid residue; and

(d) JK* is an amino acid sequence selected from the group consisting of sequences encoded by IGJK1, IGJK2, IGJK3, IGJK4, and IGJK5, wherein the first residue of each IGJK sequence is not present.

In some aspects, the VKCDR3 amino acid sequence comprises one or more of the sequences listed in Table 33 or a sequence at least about 80% identical to any of them. In other aspects, the antibody proteins are expressed in a heterodimeric form. In yet another aspect, the human antibody proteins are expressed as antibody fragments. In still other aspects of the invention, the antibody fragments are selected from the group consisting of Fab, Fab', F(ab')$_2$, Fv fragments, diabodies, linear antibodies, and single-chain antibodies.

In certain embodiments, the invention comprises an antibody isolated from the polypeptide expression products of any library described herein.

In still other aspects, the polynucleotides further comprise a 5’ polynucleotide sequence and a 3’ polynucleotide sequence that facilitate homologous recombination.

In one embodiment, the polynucleotides further encode an alternative scaffold.

In another embodiment, the invention comprises a library of polypeptides encoded by any of the synthetic polynucleotide libraries described herein.

In yet another embodiment, the invention comprises a library of vectors comprising any of the polynucleotide libraries described herein. In certain other aspects, the invention comprises a population of cells comprising the vectors of the instant invention.

In one aspect, the doubling time of the population of cells is from about 1 to about 3 hours, from about 3 to about 8 hours, from about 8 to about 16 hours, from about 16 to about 20 hours, or from 20 to about 30 hours. In yet another aspect, the cells are yeast cells. In still another aspect, the yeast is *Saccharomyces cerevisiae*.

In other embodiments, the invention comprises a library that has a theoretical total diversity of N unique CDRH3 sequences, wherein N is about 10^6 to about 10^{15}; and wherein the physical realization of the theoretical total CDRH3 diversity has a size of at least about 3N, thereby providing a probability of at least about 95% that any individual CDRH3 sequence contained within the theoretical total diversity of the library is present in the actual library.

In certain embodiments, the invention comprises a library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody VλCDR3 amino acid sequences comprising about 1 to about 10 of the amino acids found at Kabat positions 89, 90, 91, 92, 93, 94, 95, 95A, 95B, 95C, 96, and 97, in selected VλCDR3 sequences encoded by a single germline sequence.

In some embodiments, the invention relates to a library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the library has a theoretical total diversity of about 10^6 to about 10^{15} unique CDRH3 sequences.

In still other embodiments, the invention relates to a method of preparing a library of synthetic polynucleotides encoding a plurality of antibody VK amino acid sequences, the method comprising:

(i) providing polynucleotide sequences encoding:

(a) one or more VK_Chassis amino acid sequences selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27, IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30, IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-21, and IGKV6D-41, or a sequence at least about 80% identical to any of them;

(b) one or more L3-VK amino acid sequences, wherein L3-VK the portion of the VKCDR3 amino acid sequence encoded by the IGKV gene segment;

(c) one or more X residues, wherein X is any amino acid residue; and

(d) one or more JK* amino acid sequences, wherein JK* is an amino acid sequence selected from the group consisting amino acid sequences encoded by IGKJ1, IGKJ2, IGKJ3, IGKJ4, and IGKJ5, wherein the first amino acid residue of each sequence is not present; and

(ii) assembling the polynucleotide sequences to produce a library of synthetic polynucleotides encoding a plurality of human VK sequences represented by the following formula:

[VK_Chassis]-[L3-VK]-[X]-[JK*].

In some embodiments, the invention relates to a method of preparing a library of synthetic polynucleotides encoding a plurality of antibody light chain CDR3 sequences, the method comprising:

(i) determining the percent occurrence of each amino acid residue at each position in selected light chain CDR3 amino acid sequences derived from a single germline polynucleotide sequence;

(ii) designing synthetic polynucleotides encoding a plurality of human antibody light chain CDR3 amino acid sequences, wherein the percent occurrence of any amino acid at any position within the designed light chain CDR3 amino acid sequences is within at least about 30% of the percent occurrence in the selected light chain CDR3 amino acid sequences derived from a single germline polynucleotide sequence, as determined in (i); and

(iii) synthesizing one or more polynucleotides that were designed in (ii).

In other embodiments, the invention relates to a method of preparing a library of synthetic polynucleotides encoding a plurality of antibody Vλ amino acid sequences, the method comprising:

(i) providing polynucleotide sequences encoding:

(a) one or more Vλ_Chassis amino acid sequences selected from the group consisting of about Kabat residue 1 to about Kabat residue 88 encoded by IGλV1-36, IGλV1-40, IGλV1-44, IGλV1-47, IGλV1-51, IGλV10-54, IGλV2-11, IGλV2-14, IGλV2-18, IGλV2-23, IGλV2-8, IGλV3-1, IGλV3-10, IGλV3-12, IGλV3-16, IGλV3-19, IGλV3-21, IGλV3-25, IGλV3-27, IGλV3-9, IGλV4-3, IGλV4-60, IGλV4-69, IGλV5-39, IGλV5-45, IGλV6-57, IGλV7-43, IGλV7-46, IGλV8-61, IGλV9-49, and IGλV10-54, or a sequence at least about 80% identical to any of them;

(b) one ore more L3-Vλ sequences, wherein L3-Vλ is the portion of the VλCDR3 amino acid sequence encoded by the IGλV gene segment;

(c) one or more Jλ sequences, wherein Jλ is an amino acid sequence selected from the group consisting of amino acid sequences encoded by IGλJ1-01, IGλJ2-01, IGλJ3-01, IGλJ3-02, IGλJ6-01, IGλJ7-01, and IGλJ7-02 wherein the first amino acid residue of each sequence may or may not be present; and

(ii) assembling the polynucleotide sequences to produce a library of synthetic polynucleotides encoding a plurality of human Vλ amino acid sequences represented by the following formula:

[Vλ_Chassis]-[L3-Vλ]-[Jλ].

In certain embodiments, the amino acid sequences encoded by the polynucleotides of the libraries of the invention are human.

The present invention is also directed to methods of preparing a synthetic polynucleotide library comprising providing and assembling the polynucleotide sequences of the instant invention.

In another aspect, the invention comprises a method of preparing the library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, the method comprising:

(i) providing polynucleotide sequences encoding:

(a) one or more N1 amino acid sequences of about 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

(b) one or more human CDRH3 DH amino acid sequences, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(c) one or more N2 amino acid sequences of about 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

(d) one or more human CDRH3 H3-JH amino acid sequences, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them; and

(ii) assembling the polynucleotide sequences to produce a library of synthetic polynucleotides encoding a plurality of human antibody CDRH3 amino acid sequences represented by the following formula:

[N1]-[DH]-[N2]-[H3-JH].

In one aspect, one or more of the polynucleotide sequences are synthesized via split-pool synthesis.

In another aspect, the method of the invention further comprises the step of recombining the assembled synthetic polynucleotides with a vector comprising a heavy chain chassis and a heavy chain constant region, to form a full-length heavy chain.

In another aspect, the method of the invention further comprises the step of providing a 5' polynucleotide sequence and a 3' polynucleotide sequence that facilitate homologous recombination. In still another aspect, the method of the invention further comprises the step of recombining the assembled synthetic polynucleotides with a vector comprising a heavy chain chassis and a heavy chain constant region, to form a full-length heavy chain.

In some embodiments, the step of recombining is performed in yeast. In certain embodiments, the yeast is *S. cerevisiae*.

In certain other embodiments, the invention comprises a method of isolating one or more host cells expressing one or more antibodies, the method comprising:

(i) expressing the human antibodies of any one of claims 40 and 46 in one or more host cells;

(ii) contacting the host cells with one or more antigens; and

(iii) isolating one or more host cells having antibodies that bind to the one or more antigens.

In another aspect, the method of the invention further comprises the step of isolating one or more antibodies from the one or more host cells that present the antibodies which recognize the one or more antigens. In yet another aspect, the method of the invention further comprises the step of isolating one or more polynucleotide sequences encoding one or more antibodies from the one or more host cells that present the antibodies which recognize the one or more antigens.

In certain other embodiments, the invention comprises a kit comprising the library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, or any of the other sequences disclosed herein.

In still other aspects, the CDRH3 amino acid sequences encoded by the libraries of synthetic polynucleotides described herein, or any of the other sequences disclosed herein, are in computer readable form.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts a schematic of recombination between a fragment (*e.g.,* CDR3) and a vector (*e.g.,* comprising a chassis and constant region) for the construction of a library.

Figure 2 depicts the length distribution of the N1 and N2 regions of rearranged human antibody sequences compiled from Jackson *et al.* (J. Immunol Methods, 2007, 324: 26, incorporated by reference in its entirety).

Figure 3 depicts the length distribution of the CDRL3 regions of rearranged human kappa light chain sequences compiled from the NCBI database (Appendix A).

Figure 4 depicts the length distribution of the CDRL3 regions of rearranged human lambda light chain sequences compiled from the NCBI database (Appendix B).

Figure 5 depicts a schematic representation of the 424 cloning vectors used in the synthesis of the CDRH3 regions before and after ligation of the [DH]-[N2]-[JH] segment.

Figure 6 depicts a schematic structure of a heavy chain vector, prior to recombination with a CDRH3.

Figure 7 depicts a schematic diagram of a CDRH3 integrated into a heavy chain vector and the polynucleotide and polypeptide sequences of CDRH3.

Figure 8 depicts a schematic structure of a kappa light chain vector, prior to recombination with a CDRL3.

Figure 9 depicts a schematic diagram of a CDRL3 integrated into a light chain vector and the polynucleotide and polypeptide sequences of CDRL3.

Figure 10 depicts the length distribution of the CDRH3 domain (Kabat positions 95-102) from 96 colonies obtained by transformation with 10 of the 424 vectors synthesized as described in Example 10 (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 11 depicts the length distribution of the DH segment from 96 colonies obtained by transformation with 10 of the 424 vectors synthesized as described in Example 10 (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 12 depicts the length distribution of the N2 segment from 96 colonies obtained by transformation with 10 of the 424 vectors synthesized as described in Example 10 (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 13 depicts the length distribution of the H3-JH segment from 96 colonies obtained by transformation with 10 of the 424 vectors synthesized as described in Example 10 (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 14 depicts the length distribution of the CDRH3 domains from 291 sequences prepared from yeast cells transformed according to the method outlined in Example 10.4, namely the co-transformation of vectors containing heavy chain chassis and constant regions with a CDRH3 insert (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 15 depicts the length distribution of the [Tail]-[N1] region from the 291 sequences prepared from yeast cells transformed according to the protocol outlined in Example 10.4 (observed), as compared to the expected (*i.e.,* designed) distribution.

Figure 16 depicts the length distribution of the DH region from the 291 sequences prepared from yeast cells transformed according to the protocol outlined in Example 10.4 (observed), as compared to the theoretical (*i.e.,* designed) distribution.

Figure 17 depicts the length distribution of the N2 region from the 291 sequences prepared from yeast cells transformed according to the protocol outlined in Example 10.4 (observed), as compared to the theoretical (*i.e.,* designed) distribution.

Figure 18 depicts the length distribution of the H3-JH region from the 291 sequences prepared from yeast cells transformed according to the protocol outlined in Example 10.4 (observed), as compared to the theoretical (*i.e.,* designed) distribution.

Figure 19 depicts the familial origin of the JH segments identified in the 291 sequences (observed), as compared to the theoretical (*i.e.,* designed) familial origin.

Figure 20 depicts the representation of each of the 16 chassis of the library (observed), as compared to the theoretical (*i.e.,* designed) chassis representation. VH3-23 is represented twice; once ending in CAR and once ending in CAK. These representations were combined, as were the ten variants of VH3-33 with one variant of VH3-30.

Figure 21 depicts a comparison of the CDRL3 length from 86 sequences selected from the VKCDR3 library of Example 6.2 (observed) to human sequences (human) and the designed sequences (designed).

Figure 22 depicts the representation of the light chain chassis amongst the 86 sequences selected from the library (observed), as compared to the theoretical (*i.e.,* designed) chassis representation.

Figure 23 depicts the frequency of occurrence of different CDRH3 lengths in an exemplary library of the invention, versus the preimmune repertoire of Lee *et al.* (Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety).

Figure 24 depicts binding curves for 6 antibodies selected from a library of the invention.

Figure 25 depicts binding curves for 10 antibodies selected from a library of the invention binding to hen egg white lysozyme.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to, at least, synthetic polynucleotide libraries, methods of producing and using the libraries of the invention, kits and computer readable forms including the libraries of the invention. The libraries taught in this application are described, at least in part, in terms of the components from which they are assembled.

In certain embodiments, the instant invention provides antibody libraries specifically designed based on the composition and CDR length distribution in the naturally occurring human antibody repertoire. It is estimated that, even in the absence of antigenic stimulation, a human makes at least about 10^7 different antibody molecules. The antigen-binding sites of many antibodies can cross-react with a variety of related but different epitopes. In addition the human antibody repertoire is large enough to ensure that there is an antigen-binding site to fit almost any potential epitope, albeit with low affinity.

The mammalian immune system has evolved unique genetic mechanisms that enable it to generate an almost unlimited number of different light and heavy chains in a remarkably economical way, by combinatorially joining chromosomally separated gene segments prior to transcription. Each type of immunoglobulin (Ig) chain (*i.e.,* κ light, λ light, and heavy) is synthesized by combinatorial assembly of DNA sequences selected from two or more families of gene segments, to produce a single polypeptide chain. Specifically, the heavy chains and light chains each consist of a variable region and a constant (C) region. The variable regions of the heavy chains are encoded by DNA sequences assembled from three families of gene segments: variable (IGHV), joining (IGHJ) and diversity (IGHD). The variable regions of light chains are encoded by DNA sequences assembled from two families of gene segments for each of the kappa and lambda light chains: variable (IGLV) and joining (IGLJ). Each variable region (heavy and light) is also recombined with a constant region, to produce a full-length immunoglobulin chain.

While combinatorial assembly of the V, D and J gene segments make a substantial contribution to antibody variable region diversity, further diversity is introduced *in vivo*, at the pre-B cell stage, via imprecise joining of these gene segments and the introduction of non-templated nucleotides at the junctions between the gene segments.

After a B cell recognizes an antigen, it is induced to proliferate. During proliferation, the B cell receptor locus undergoes an extremely high rate of somatic mutation that is far greater than the normal rate of genomic mutation. The mutations that occur are primarily localized to the Ig variable regions and comprise substitutions, insertions and deletions. This somatic hypermutation enables the production of B cells that express antibodies possessing enhanced affinity toward an antigen. Such antigen-driven somatic hypermutation fine-tunes antibody responses to a given antigen.

Significant efforts have been made to create antibody libraries with extensive diversity, and to mimic the natural process of affinity maturation of antibodies against various antigens, especially antigens associated with diseases such as autoimmune diseases, cancer, and infectious disease. Antibody libraries comprising candidate binding molecules that can be readily screened against targets are desirable. However, the full promise of an antibody library, which is representative of the preimmune human antibody repertoire, has remained elusive. In addition to the shortcomings enumerated above, and throughout the application, synthetic libraries that are known in the art often suffer from noise (*i.e.,* very large libraries increase the presence of many sequences which do not express well, and/or which misfold), while entirely human libraries that are known in the art may be biased against certain antigen classes (*e.g.,* self-antigens). Moreover, the limitations of synthesis and physical realization techniques restrict the functional diversity of antibody libraries of the art. The present invention provides, for the first time, a fully synthetic antibody library that is representative of the human preimmune antibody repertoire (*e.g.,* in composition and length), and that can be readily screened (*i.e.,* it is physically realizable and, in some cases can be oversampled) using, for example, high throughput methods, to obtain, for example, new therapeutics and/or diagnostics

In particular, the synthetic antibody libraries of the instant invention have the potential to recognize any antigen, including self-antigens of human origin. The ability to recognize self-antigens is usually lost in an expressed human library, because self-reactive antibodies are removed by the donor's immune system via negative selection. Another feature of the invention is that screening the antibody library using positive clone selection, for example, by FACS (florescence activated cell sorter) bypasses the standard and tedious methodology of generating a hybridoma library and supernatant screening. Still further, the libraries, or sub-libraries thereof, can be screened multiple times, to discover additional antibodies against other desired targets.

Before further description of the invention, certain terms are defined.

1. Definitions

Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by one of ordinary skill in the art relevant to the invention. The definitions below supplement those in the art and are directed to the embodiments described in the current application.

The term "antibody" is used herein in the broadest sense and specifically encompasses at least monoclonal antibodies, polyclonal antibodies, multi-specific antibodies (*e.g.,* bispecific antibodies), chimeric antibodies, humanized antibodies, human antibodies, and antibody fragments. An antibody is a protein comprising one or more polypeptides substantially or partially encoded by immunoglobulin genes or fragments of immunoglobulin genes. The recognized immunoglobulin genes include the kappa, lambda, alpha, gamma, delta, epsilon and mu constant region genes, as well as myriad immunoglobulin variable region genes.

"Antibody fragments" comprise a portion of an intact antibody, for example, one or more portions of the antigen-binding region thereof. Examples of antibody fragments include Fab, Fab', $F(ab')_2$, and Fv fragments, diabodies, linear antibodies, single-chain antibodies, and multi-specific antibodies formed from intact antibodies and antibody fragments.

An "intact antibody" is one comprising full-length heavy- and light- chains and an Fc region. An intact antibody is also referred to as a “full-length, heterodimeric” antibody or immunoglobulin.

The term "variable" refers to the portions of the immunoglobulin domains that exhibit variability in their sequence and that are involved in determining the specificity and binding affinity of a particular antibody (*i.e.,* the “variable domain(s)”). Variability is not evenly distributed throughout the variable domains of antibodies; it is concentrated in sub-domains of each of the heavy and light chain variable regions. These sub-domains are called “hypervariable” regions or “complementarity determining regions” (CDRs). The more conserved (*i.e.,* non-hypervariable) portions of the variable domains are called the “framework” regions (FRM). The variable domains of naturally occurring heavy and light chains each comprise four FRM regions, largely adopting a β-sheet configuration, connected by three hypervariable regions, which form loops connecting, and in some cases forming part of, the β -sheet structure. The hypervariable regions in each chain are held together in close proximity by the FRM and, with the hypervariable regions from the other chain, contribute to the formation of the antigen-binding site (see Kabat *et al.* Sequences of Proteins of Immunological Interest, 5th Ed. Public Health Service, National Institutes of Health, Bethesda, Md., 1991, incorporated by reference in its entirety). The constant domains are not directly involved in antigen binding, but exhibit various effector functions, such as, for example, antibody-dependent, cell-mediated cytotoxicity and complement activation.

The "chassis" of the invention represent a portion of the antibody heavy chain variable (IGHV) or light chain variable (IGLV) domains that are not part of CDRH3 or CDRL3, respectively. The chassis of the invention is defined as the portion of the variable region of an antibody beginning with the first amino acid of FRM1 and ending with the last amino acid of FRM3. In the case of the heavy chain, the chassis includes the amino acids including from about Kabat position 1 to about Kabat position 94. In the case of the light chains (kappa and lambda), the chassis are defined as including from about Kabat position 1 to about Kabat position 88. The chassis of the invention may contain certain modifications relative to the corresponding germline variable domain sequences presented herein or available in public databases. These modifications may be engineered (*e.g.,* to remove N-linked glycosylation sites) or naturally occurring (*e.g.,* to account for allelic variation). For example, it is known in the art that the immunoglobulin gene repertoire is polymorphic (Wang *et al.,* Immunol. Cell. Biol., 2008, 86: 111; Collins *et al.,* Immunogenetics, 2008, DOI 10.1007/s00251-008-0325-z, published online, each incorporated by reference in its entirety); chassis, CDRs (*e.g.,* CDRH3) and constant regions representative of these allelic variants are also encompassed by the invention. In some embodiments, the allelic variant(s) used in a particular embodiment of the invention may be selected based on the allelic variation present in different patient populations, for example, to identify antibodies that are non-immunogenic in these patient populations. In certain embodiments, the immunogenicity of an antibody of the invention may depend on allelic variation in the major histocompatibility complex (MHC) genes of a patient population. Such allelic variation may also be considered in the design of libraries of the invention. In certain embodiments of the invention, the chassis and constant regions are contained on a vector, and a CDR3 region is introduced between them via homologous recombination.

In some embodiments, one, two or three nucleotides may follow the heavy chain chassis, forming either a partial (if one or two) or a complete (if three) codon. When a full codon is present, these nucleotides encode an amino acid residue that is referred to as the "tail," and occupies position 95.

The "CDRH3 numbering system" used herein defines the first amino acid of CDRH3 as being at Kabat position 95 (the "tail," when present) and the last amino acid of CDRH3 as position 102. The amino acids following the "tail" are called "N1" and, when present, are assigned numbers 96, 96A, 96B, etc. The N1 segment is followed by the "DH" segment, which is assigned numbers 97, 97A, 97B, 97C, etc. The DH segment is followed by the "N2" segment, which, when present, is numbered 98, 98A, 98B, etc. Finally, the most C-terminal amino acid residue of the set of the "H3-JH" segment is designated as number 102. The residue directly before (N-terminal) it, when present, is 101, and the one before (if present) is 100. For reasons of convenience, and which will become apparent elsewhere, the rest of the H3-JH amino acids are numbered in reverse order, beginning with 99 for the amino acid just N-terminal to 100, 99A for the residue N-terminal to 99, and so forth for 99B, 99C, etc. Examples of certain CDRH3 sequence residue numbers may therefore include the following:

13 Amino Acid CDR-H3 with N1 and N2

```
(95)(96)(96A)(97)(97A)(97B)(97C)(97D) (98) (99)(100)(101)(102)
|---|--------|----------------------|----|------------------|
 Tail   N1              DH              N2        H3-JH
```

10 Amino Acid CDR-H3 without N1 and N2

```
(97)(97A)(97B)(97C)(97D)(97E)(97F)(97G)(101)(102)
|------------------------------------|---------|
                  DH                    H3-JH
```

As used herein, the term "diversity" refers to a variety or a noticeable heterogeneity. The term "sequence diversity" refers to a variety of sequences which are collectively representative of several possibilities of sequences, for example, those found in natural human antibodies. For example, heavy chain CDR3 (CDRH3) sequence diversity may refer to a variety of possibilities of combining the known human DH and H3-JH segments, including the N1 and N2 regions, to form heavy chain CDR3 sequences. The light chain CDR3 (CDRL3) sequence diversity may refer to a variety of possibilities of combining the naturally occurring light chain variable region contributing to CDRL3 (*i.e.,* L3-VL) and joining (*i.e.,* L3-JL) segments, to form light chain CDR3 sequences. As used herein, H3-JH refers to the portion of the IGHJ gene contributing to CDRH3. As used herein, L3-VL and L3-JL refer to the portions of the IGLV and IGLJ genes (kappa or lambda) contributing to CDRL3, respectively.

As used herein, the term “expression” includes any step involved in the production of a polypeptide including, but not limited to, transcription, post-transcriptional modification, translation, post-translational modification, and secretion.

As used herein, the term "host cell" is intended to refer to a cell into which a polynucleotide of the invention. It should be understood that such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

The term “length diversity” refers to a variety in the length of a particular nucleotide or amino acid sequence. For example, in naturally occurring human antibodies, the heavy chain CDR3 sequence varies in length, for example, from about 3 amino acids to over about 35 amino acids, and the light chain CDR3 sequence varies in length, for example, from about 5 to about 16 amino acids. Prior to the instant invention, it was known in the art that it is possible to design antibody libraries containing sequence diversity *or* length diversity (see, *e.g.,* Hoet *et al.,* Nat. Biotechnol., 2005, 23: 344; Kretzschmar and von Ruden, Curr. Opin. Biotechnol., 2002 13: 598; and Rauchenberger *et al.,* J. Biol. Chem., 2003 278: 38194, each of which is incorporated by reference in its entirety); however, the instant invention is directed to, at least, the design of synthetic antibody libraries containing the sequence diversity *and* length diversity of naturally occurring human sequences. In some cases, synthetic libraries containing sequence and length diversity have been synthesized, however these libraries contain too much theoretical diversity to synthesize the entire designed repertoire and/or too many theoretical members to physically realize or oversample the entire library.

As used herein, a sequence designed with “directed diversity” has been specifically designed to contain both sequence diversity and length diversity. Directed diversity is not stochastic.

As used herein, “stochastic” describes a process of generating a randomly determined sequence of amino acids, which is considered as a sample of one element from a probability distribution.

The term “library of polynucleotides” refers to two or more polynucleotides having a diversity as described herein, specifically designed according to the methods of the invention. The term “library of polypeptides” refers to two or more polypeptides having a diversity as described herein, specifically designed according to the methods of the invention. The term "library of synthetic polynucleotides" refers to a polynucleotide library that includes synthetic polynucleotides. The term "library of vectors" refers herein to a library of at least two different vectors. As used herein, the term “human antibody libraries,” at least includes, a polynucleotide or polypeptide library which has been designed to represent the sequence diversity and length diversity of naturally occurring human antibodies.

As described throughout the specification, the term “library” is used herein in its broadest sense, and also may include the sub-libraries that may or may not be combined to produce libraries of the invention.

As used herein, the term “synthetic polynucleotide” refers to a molecule formed through a chemical process, as opposed to molecules of natural origin, or molecules derived via template-based amplification of molecules of natural origin (*e.g.,* immunoglobulin chains cloned from populations of B cells via PCR amplification are not “synthetic” used herein). In some instances, for example, when referring to libraries of the invention that comprise multiple components (*e.g.,* N1, DH, N2, and/or H3-JH), the invention encompasses libraries in which at least one of the aforementioned components is synthetic. By way of illustration, a library in which certain components are synthetic, while other components are of natural origin or derived via template-based amplification of molecules of natural origin, would be encompassed by the invention.

The term "split-pool synthesis" refers to a procedure in which the products of a plurality of first reactions are combined (pooled) and then separated (split) before participating in a plurality of second reactions. Example 9, describes the synthesis of 278 DH segments (products), each in a separate reaction. After synthesis, these 278 segments are combined (pooled) and then distributed (split) amongst 141 columns for the synthesis of the N2 segments. This enables the pairing of each of the 278 DH segments with each of the 141 N2 segments. As described elsewhere in the specification, these numbers are non-limiting.

“Preimmune” antibody libraries have similar sequence diversities and length diversities to naturally occurring human antibody sequences before these sequences have undergone negative selection or somatic hypermutation. For example, the set of sequences described in Lee *et al.* (Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety) is believed to represent sequences from the preimmune repertoire. In certain embodiments of the invention, the sequences of the invention will be similar to these sequences (*e.g.,* in terms of composition and length). In certain embodiments of the invention, such antibody libraries are designed to be small enough to chemically synthesize and physically realize, but large enough to encode antibodies with the potential to recognize any antigen. In one embodiment of the invention, an antibody library comprises about 10^7 to about 10^{20} different antibodies and/or polynucleotide sequences encoding the antibodies of the library. In some embodiments, the libraries of the instant invention are designed to include 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, or 10^{20} different antibodies and/or polynucleotide sequences encoding the antibodies. In certain embodiments, the libraries of the invention may comprise or encode about 10^3 to about 10^5, about 10^5 to about 10^7, about 10^7 to about 10^9, about 10^9 to about 10^{11}, about 10^{11} to about 10^{13}, about 10^{13} to about 10^{15}, about 10^{15} to about 10^{17}, or about 10^{17} to about 10^{20} different antibodies. In certain embodiments of the invention, the diversity of the libraries may be characterized as being greater than or less than one or more of the diversities enumerated above, for example greater than about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, or 10^{20} or less than about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, or 10^{20}. In certain other embodiments of the invention, the probability of an antibody of interest being present in a physical realization of a library with a size as enumerated above is at least about 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 85%, 90%, 95%, 99%, 99.5%, or 99.9% (see Library Sampling, in the Detailed Description, for more information on the probability of a particular sequence being present in a physical realization of a library). The antibody libraries of the invention may also include antibodies directed to, for example, self (*i.e.,* human) antigens. The antibodies of the present invention may not be present in expressed human libraries for reasons including because self-reactive antibodies are removed by the donor's immune system via negative selection. However, novel heavy/light chain pairings may in some cases create self-reactive antibody specificity (Griffiths et al. US Patent 5,885,793, incorporated by reference in its entirety). In certain embodiments of the invention, the number of unique heavy chains in a library may be about 10, 50, 10^2, 150, 10^3, 10^4, 10^5,10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18},

10^{19}, 10^{20}, or more. In certain embodiments of the invention, the number of unique light chains in a library may be about 5, 10, 25, 50, 10^2, 150, 500, 10^3, 10^4, 10^5,10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, 10^{20}, or more.

As used herein, the term "human antibody CDRH3 libraries," at least includes, a polynucleotide or polypeptide library which has been designed to represent the sequence diversity and length diversity of naturally occurring human antibodies. "Preimmune" CDRH3 libraries have similar sequence diversities and length diversities to naturally occurring human antibody CDRH3 sequences before these sequences undergo negative selection and somatic hypermutation. Known human CDRH3 sequences are represented in various data sets, including Jackson *et al.,* J. Immunol Methods, 2007, 324: 26; Martin, Proteins, 1996, 25: 130; and Lee *et al.,* Immunogenetics, 2006, 57: 917, each of which is incorporated by reference in its entirety. In certain embodiments of the invention, such CDRH3 libraries are designed to be small enough to chemically synthesize and physically realize, but large enough to encode CDRH3s with the potential to recognize any antigen. In one embodiment of the invention, an antibody library includes about 10^6 to about 10^{15} different CDRH3 sequences and/or polynucleotide sequences encoding said CDRH3 sequences. In some embodiments, the libraries of the instant invention are designed to about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, or 10^{16}, different CDRH3 sequences and/or polynucleotide sequences encoding said CDRH3 sequences. In some embodiments, the libraries of the invention may include or encode about10^3 to about 10^6, about 10^6 to about 10^8, about 10^8 to about 10^{10}, about 10^{10} to about 10^{12}, about 10^{12} to about 10^{14}, or about 10^{14} to about 10^{16} different CDRH3 sequences. In certain embodiments of the invention, the diversity of the libraries may be characterized as being greater than or less than one or more of the diversities enumerated above, for example greater than about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, or 10^{16} or less than about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, or 10^{16}. In certain embodiments of the invention, the probability of a CDRH3 of interest being present in a physical realization of a library with a size as enumerated above is at least about 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 85%, 90%, 95%, 99%, 99.5%, or 99.9% (see Library Sampling, in the Detailed Description, for more information on the probability of a particular sequence being present in a physical realization of a library).

The preimmune CDRH3 libraries of the invention may also include CDRH3s directed to, for example, self (*i.e.,* human) antigens. Such CDRH3s may not be present in expressed human libraries, because self-reactive CDRH3s are removed by the donor's immune system via negative selection.

Libraries of the invention containing "VKCDR3" sequences and "VλCDR3" sequences refer to the kappa and lambda sub-sets of the CDRL3 sequences, respectively. These libraries may be designed with directed diversity, to collectively represent the length and sequence diversity of the human antibody CDRL3 repertoire. "Preimmune" versions of these libraries have similar sequence diversities and length diversities to naturally occurring human antibody CDRL3 sequences before these sequences undergo negative selection. Known human CDRL3 sequences are represented in various data sets, including the NCBI database (see Appendix A and Appendix B for light chain sequence data sets) and Martin, Proteins, 1996, 25: 130 incorporated by reference in its entirety. In certain embodiments of the invention, such CDRL3 libraries are designed to be small enough to chemically synthesize and physically realize, but large enough to encode CDRL3s with the potential to recognize any antigen.

In one embodiment of the invention, an antibody library comprises about 10^5 different CDRL3 sequences and/or polynucleotide sequences encoding said CDRL3 sequences. In some embodiments, the libraries of the instant invention are designed to comprise about 10^1, 10^2, 10^3, 10^4, 10^6, 10^7, or 10^8 different CDRL3 sequences and/or polynucleotide sequences encoding said CDRL3 sequences. In some embodiments, the libraries of the invention may comprise or encode about 10^1 to about 10^3, about 10^3 to about 10^5, or about 10^5 to about 10^8 different CDRL3 sequences. In certain embodiments of the invention, the diversity of the libraries may be characterized as being greater than or less than one or more of the diversities enumerated above, for example greater than about 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, or 10^8 or less than about 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, or 10^8. In certain embodiments of the invention, the probability of a CDRL3 of interest being present in a physical realization of a library with a size as enumerated above is at least about 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 85%, 90%, 95%, 99%, 99.5%, or 99.9% (see Library Sampling, in the Detailed Description, for more information on the probability of a particular sequence being present in a physical realization of a library). The preimmune CDRL3 libraries of the invention may also include CDRL3s directed to, for example, self (*i.e.,* human) antigens. Such CDRL3s may not be present in expressed human libraries, because self-reactive CDRL3s are removed by the donor's immune system via negative selection.

As used herein, the term "known heavy chain CDR3 sequences" refers to heavy chain CDR3 sequences in the public domain that have been cloned from populations of human B cells. Examples of such sequences are those published or derived from public data sets, including, for example, Zemlin *et al.,* JMB, 2003, 334: 733; Lee *et al.,* Immunogenetics, 2006, 57: 917; and Jackson *et al.* J. Immunol Methods, 2007, 324: 26, each of which are incorporated by reference in their entirety.

As used herein, the term "known light chain CDR3 sequences" refers to light chain CDR3 sequences (*e.g.,* kappa or lambda) in the public domain that have been cloned from populations of human B cells. Examples of such sequences are those published or derived from public data sets, including, for example, the NCBI database (see Appendices A and B filed herewith).

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 (or variable region or one or more CDRs thereof), an antibody heavy chain (or variable region or one or more CDRs thereof), a heavy chain Fd region, a combined antibody light and heavy chain (or variable regions thereof) such as a Fab, $F(ab')_2$, single domain, or single chain antibodies (scFv), or any region of a full length antibody that recognizes an antigen, 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 (*i.e.,* hypervariable) CDRs. Such framework regions are typically referred to as frameworks 1 through 4 (FRM1, FRM2, FRM3, and FRM4) and provide a scaffold for the presentation of the six CDRs (three from the heavy chain and three from the light chain) in three dimensional space, to form an antigen-binding surface.

The term "canonical structure" refers to the main chain conformation that is adopted by the antigen binding (CDR) loops. From comparative structural studies, it has been found that five of the six antigen binding loops have only a limited repertoire of available conformations. Each canonical structure can be characterized by the torsion angles of the polypeptide backbone. 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, J. Mol. Biol., 1987, 196: 901; Chothia *et al.*, Nature, 1989, 342: 877; Martin and Thornton, J. Mol. Biol., 1996, 263: 800, each of which is incorporated by reference in its entirety). Furthermore, there is a relationship between the adopted loop structure and the amino acid sequences surrounding it. The conformation of a particular canonical class is determined by the length of the loop and the amino acid residues residing at key positions within the loop, as well as within the conserved framework (*i.e.,* outside of the loop). Assignment to a particular canonical class can therefore be made based on the presence of these key amino acid residues. The term "canonical structure" may also include considerations as to the linear sequence of the antibody, for example, as catalogued by Kabat (Kabat *et al.,* in "Sequences of Proteins of Immunological Interest," 5th Edition, U.S. Department of Heath and Human Services, 1992). 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 can also be used to determine the canonical structure of an antibody. For example, those differences not fully reflected by Kabat numbering can be described by the numbering system of Chothia *et al.* and/or revealed by other techniques, for example, crystallography and two or three-dimensional computational modeling. Accordingly, a given antibody sequence may be placed into a canonical class which allows for, among other things, identifying appropriate chassis sequences (*e.g.,* based on a desire to include a variety of canonical structures in a library). Kabat numbering of antibody amino acid sequences and structural considerations as described by Chothia *et al.*, and their implications for construing canonical aspects of antibody structure, are described in the literature.

The terms "CDR", and its plural "CDRs", refer to a complementarity determining region (CDR) of which three make up the binding character of a light chain variable region (CDRL1, CDRL2 and CDRL3) and three make up the binding character of a heavy chain variable region (CDRH1, CDRH2 and CDRH3). CDRs contribute to the functional activity of an antibody molecule and are separated by amino acid sequences that comprise scaffolding or framework regions. The exact definitional CDR boundaries and lengths are subject to different classification and numbering systems. CDRs may therefore be referred to by Kabat, Chothia, contact or any other boundary definitions, including the numbering system described herein. Despite differing boundaries, each of these systems has some degree of overlap in what constitutes the so called "hypervariable regions" within the variable sequences. CDR definitions according to these systems may therefore differ in length and boundary areas with respect to the adjacent framework region. See for example Kabat, Chothia, and/or MacCallum *et al.*, (Kabat *et al.*, in “Sequences of Proteins of Immunological Interest,” 5th Edition, U.S. Department of Health and Human Services, 1992; Chothia *et al.*, J. Mol. Biol., 1987, 196: 901; and MacCallum *et al.*, J. Mol. Biol., 1996, 262: 732, each of which is incorporated by reference in its entirety).

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 or A); arginine (Arg or R); asparagine (Asn or N); aspartic acid (Asp or D); cysteine (Cys or C); glutamine (Gln or Q); glutamic acid (Glu or E); glycine (Gly or G); histidine (His or H); isoleucine (Ile or I): leucine (Leu or L); lysine (Lys or K); methionine (Met or M); phenylalanine (Phe or F); proline (Pro or P); serine (Ser or S); threonine (Thr or T); tryptophan (Trp or W); tyrosine (Tyr or Y); and valine (Val or V), although modified, synthetic, or rare amino acids may be used as desired. Generally, amino acids can be grouped as having a nonpolar side chain (*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 side chain (*e.g.*, Asn, Cys, Gln, Gly, His, Met, Phe, Ser, Thr, Trp, and Tyr).

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 terms “theoretical diversity”, “theoretical total diversity”, or “theoretical repertoire” refer to the maximum number of variants in a library design. For example, given an amino acid sequence of three residues, where residues one and three may each be any one of five amino acid types and residue two may be any one of 20 amino acid types, the theoretical diversity is 5×20×5=500 possible sequences. Similarly if sequence X is constructed by combination of 4 amino acid segments, where segment 1 has 100 possible sequences, segment 2 has 75 possible sequences, segment 3 has 250 possible sequences, and segment 4 has 30 possible sequences, the theoretical total diversity of fragment X would be 100×75×200×30, or 5.6×10^5 possible sequences.

The term "physical realization" refers to a portion of the theoretical diversity that can actually be physically sampled, for example, by any display methodology. Exemplary display methodology include: phage display, ribosomal display, and yeast display. For synthetic sequences, the size of the physical realization of a library depends on (1) the fraction of the theoretical diversity that can actually be synthesized, and (2) the limitations of the particular screening method. Exemplary limitations of screening methods include the number of variants that can be screened in a particular assay (*e.g.,* ribosome display, phage display, yeast display) and the transformation efficiency of a host cell (*e.g.,* yeast, mammalian cells, bacteria) which is used in a screening assay. For the purposes of illustration, given a library with a theoretical diversity of 10^{12} members, an exemplary physical realization of the library (*e.g.,* in yeast, bacterial cells, ribosome display, etc.; details provided below) that can maximally include 10^{11} members will, therefore, sample about 10% of the theoretical diversity of the library. However, if less than 10^{11} members of the library with a theoretical diversity of 10^{12} are synthesized, and the physical realization of the library can maximally include 10^{11} members, less than 10% of the theoretical diversity of the library is sampled in the physical realization of the library. Similarly, a physical realization of the library that can maximally include more than 10^{12} members would "oversample" the theoretical diversity, meaning that each member may be present more than once (assuming that the entire 10^{12} theoretical diversity is synthesized).

The term "all possible reading frames" encompasses at least the three forward reading frames and, in some embodiments, the three reverse reading frames.

The term "antibody of interest" refers to any antibody that has a property of interest that is isolated from a library of the invention. The property of interest may include, but is not limited to, binding to a particular antigen or epitope, blocking a binding interaction between two molecules, or eliciting a certain biological effect.

The term "functionally expressed" refers to those immunoglobulin genes that are expressed by human B cells and that do not contain premature stop codons.

The term “full-length heavy chain” refers to an immunoglobulin heavy chain that contains each of the canonical structural domains of an immunoglobulin heavy chain, including the four framework regions, the three CDRs, and the constant region. The term “full-length light chain” refers to an immunoglobulin light chain that contains each of the canonical structural domains of an immunoglobulin light chain, including the four framework regions, the three CDRs, and the constant region.

The term “unique,” as used herein, refers to a sequence that is different (e.g. has a different chemical structure) from every other sequence within the designed theoretical diversity. It should be understood that there are likely to be more than one copy of many unique sequences from the theoretical diversity in a particular physical realization. For example, a library comprising three unique sequences may comprise nine total members if each sequence occurs three times in the library. However, in certain embodiments, each unique sequence may occur only once.

The term “heterologous moiety” is used herein to indicate the addition of a composition to an antibody wherein the composition is not normally part of the antibody. Exemplary heterologous moieties include drugs, toxins, imaging agents, and any other compositions which might provide an activity that is not inherent in the antibody itself.

As used herein, the term “percent occurrence of each amino acid residue at each position” refers to the percentage of instances in a sample in which an amino acid is found at a defined position within a particular sequence. For example, given the following three sequences:

K V R
K Y P
K R P,

K occurs in position one in 100% of the instances and P occurs in position three in about 67% of the instances. In certain embodiments of the invention, the sequences selected for comparison are human immunoglobulin sequences.

As used herein, the term “most frequently occurring amino acids“ at a specified position of a sequence in a population of polypeptides refers to the amino acid residues that have the highest percent occurrence at the indicated position in the indicated polypeptide population. For example, the most frequently occurring amino acids in each of the three most N-terminal positions in N1 sequences of CDRH3 sequences that are functionally expressed by human B cells are listed in Table 21, and the most frequently occurring amino acids in each of the three most N-terminal positions in N2 sequences of CDRH3 sequences that are functionally expressed by human B cells are listed in Table 22.

For the purposes of analyzing the occurrence of certain duplets (Example 13) and the information content (Example 14) of the libraries of the invention, and other libraries, a "central loop" of CDRH3 is defined. If the C-terminal 5 amino acids from Kabat CDRH3 (95-102) are removed, then the remaining sequence is termed the "central loop". Thus, considering the duplet occurrence calculations of Example 13, using a CDRH3 of size 6 or less would not contribute to the analysis of the occurrence of duplets. A CDRH3 of size 7 would contribute only to the i – i+1 data set, a CDRH3 of size 8 would also contribute to the i – i+2 data set, and a CDRH3 of size 9 and larger would also contribute to the i – i+3 data set. For example, a CDR H3 of size 9 may have amino acids at positions 95-96-97-98-99-100-100A-101-102, but only the first four residues (bolded) would be part of the central loop and contribute to the pair-wise occurrence (duplet) statistics. As a further example, a CDRH3 of size 14 may have the sequence: 95-96-97-98-99-100-100A-100B-100C-100D-100E-100F-101-102. Here, only the first nine residues (bolded) contribute to the central loop.

Library screening requires a genotype-phenotype linkage. The term "genotype-phenotype linkage" is used in a manner consistent with its art-recognized meaning and refers to the fact that the nucleic acid (genotype) encoding a protein with a particular phenotype (*e.g.,* binding an antigen) can be isolated from a library. For the purposes of illustration, an antibody fragment expressed on the surface of a phage can be isolated based on its binding to an antigen (*e.g.,* Ladner et al.). The binding of the antibody to the antigen simultaneously enables the isolation of the phage containing the nucleic acid encoding the antibody fragment. Thus, the phenotype (antigen-binding characteristics of the antibody fragment) has been "linked" to the genotype (nucleic acid encoding the antibody fragment). Other methods of maintaining a genotype-phenotype linkage include those of Wittrup *et al.* (US Patent Nos. 6,300,065, 6,331,391, 6,423,538, 6,696,251, 6,699,658, and US Pub. No. 20040146976, each of which is incorporated by reference in its entirety), Miltenyi (US Patent No. 7,166,423, incorporated by reference in its entirety), Fandl (US Patent No. 6,919,183, US Pub No. 20060234311, each incorporated by reference in its entirety), Clausell-Tormos *et al.* (Chem. Biol., 2008, 15:

427, incorporated by reference in its entirety), Love *et al.* (Nat. Biotechnol., 2006, 24: 703, incorporated by reference in its entirety), and Kelly *et al.* (Chem. Commun., 2007, 14: 1773, incorporated by reference in its entirety). Any method which localizes the antibody protein with the gene encoding the antibody, in a way in which they can both be recovered while the linkage between them is maintained, is suitable.

2. Design of the Libraries

The antibody libraries of the invention are designed to reflect certain aspects of the preimmune repertoire as naturally created by the human immune system. Certain libraries of the invention are based on rational design informed by the collection of human V, D, and J genes, and other large databases of human heavy and light chain sequences (*e.g.,* publicly known germline sequences; sequences from Jackson *et al.,* J. Immunol Methods, 2007, 324: 26, incorporated by reference in its entirety; sequences from Lee *et al.,* Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety; and sequences compiled for rearranged VK and Vλ – see Appendices A and B filed herewith). Additional information may be found, for example, in Scaviner *et al.,* Exp. Clin. Immunogenet., 1999, 16: 234; Tomlinson *et al.,* J. Mol. Biol., 1992, 227: 799; and Matsuda *et al.,* J. Exp. Med., 1998, 188: 2151 each incorporated by reference in its entirety. In certain embodiments of the invention, cassettes representing the possible V, D, and J diversity found in the human repertoire, as well as junctional diversity (*i.e.,* N1 and N2), are synthesized *de novo* as single or double-stranded DNA oligonucleotides. In certain embodiments of the invention, oligonucleotide cassettes encoding CDR sequences are introduced into yeast along with one or more acceptor vectors containing heavy or light chain chassis sequences. No primer-based PCR amplification or template-directed cloning steps from mammalian cDNA or mRNA are employed. Through standard homologous recombination, the recipient yeast recombines the cassettes (*e.g.,* CDR3s) with the acceptor vector(s) containing the chassis sequence(s) and constant regions, to create a properly ordered synthetic, full-length human heavy chain and/or light chain immunoglobulin library that can be genetically propagated, expressed, displayed, and screened. One of ordinary skill in the art will readily recognize that the chassis contained in the acceptor vector can be designed so as to produce constructs other than full-length human heavy chains and/or light chains. For example, in certain embodiments of the invention, the chassis may be designed to encode portions of a polypeptide encoding an antibody fragment or subunit of an antibody fragment, so that a sequence encoding an antibody fragment, or subunit thereof, is produced when the oligonucleotide cassette containing the CDR is recombined with the acceptor vector.

In certain embodiments, the invention provides a synthetic, preimmune human antibody repertoire comprising about 10^7 to about 10^{20} antibody members, wherein the repertoire comprises:

(a) selected human antibody heavy chain chassis (*i.e.,* amino acids 1 to 94 of the heavy chain variable region, using Kabat's definition);

(b) a CDRH3 repertoire, designed based on the human IGHD and IGHJ germline sequences, the CDRH3 repertoire comprising the following:

(i) optionally, one or more tail regions;

(ii) one or more N1 regions, comprising about 0 to about 10 amino acids selected from the group consisting of fewer than 20 of the amino acid types preferentially encoded by the action of terminal deoxynucleotidyl transferase (TdT) and functionally expressed by human B cells;

(iii) one or DH segments, based on one or more selected IGHD segments, and one or more N- or C-terminal truncations thereof;

(iv) one or more N2 regions, comprising about 0 to about 10 amino acids selected from the group consisting of fewer than 20 of the amino acids preferentially encoded by the activity of TdT and functionally expressed by human B cells; and

(v) one or more H3-JH segments, based on one or more IGHJ segments, and one or more N-terminal truncations thereof (*e.g.,* down to XXWG);

(c) one or more selected human antibody kappa and/or lambda light chain chassis; and

(d) a CDRL3 repertoire designed based on the human IGLV and IGLJ germline sequences, wherein "L" may be a kappa or lambda light chain.

The heavy chain chassis may be any sequence with homology to Kabat residues 1 to 94 of an immunoglobulin heavy chain variable domain. Non-limiting examples of heavy chain chassis are included in the Examples, and one of ordinary skill in the art will readily recognize that the principles presented therein, and throughout the specification, may be used to derive additional heavy chain chassis.

As described above, the heavy chain chassis region is followed, optionally, by a "tail" region. The tail region comprises zero, one, or more amino acids that may or may not be selected on the basis of comparing naturally occurring heavy chain sequences. For example, in certain embodiments of the invention, heavy chain sequences available in the art may be compared, and the residues occurring most frequently in the tail position in the naturally occurring sequences included in the library (*e.g.,* to produce sequences that most closely resemble human sequences). In other embodiments, amino acids that are used less frequently may be used. In still other embodiments, amino acids selected from any group of amino acids may be used. In certain embodiments of the invention, the length of the tail is zero (no residue) or one (*e.g.,* G/D/E) amino acid. For the purposes of clarity, and without being bound by theory, in the naturally occurring human repertoire, the first 2/3 of the codon encoding the tail residue is provided by the FRM3 region of the VH gene. The amino acid at this position in naturally occurring heavy chain sequences may thus be considered to be partially encoded by the IGHV gene (2/3) and partially encoded by the CDRH3 (1/3). However, for the purposes of clearly illustrating certain aspects of the invention, the entire codon encoding the tail residue (and, therefore, the amino acid derived from it) is described herein as being part of the CDRH3 sequence.

As described above, there are two peptide segments derived from nucleotides which are added by TdT in the naturally occurring human antibody repertoire. These segments are designated N1 and N2 (referred to herein as N1 and N2 segments, domains, regions or sequences). In certain embodiments of the invention, N1 and N2 are about 0, 1, 2, or 3 amino acids in length. Without being bound by theory, it is thought that these lengths most closely mimic the N1 and N2 lengths found in the human repertoire (see Figure 2). In other embodiments of the invention, N1 and N2 may be about 4, 5, 6, 7, 8, 9, or 10 amino acids in length. Similarly, the composition of the amino acid residues utilized to produce the N1 and N2 segments may also vary. In certain embodiments of the invention, the amino acids used to produce N1 and N2 segments may be selected from amongst the eight most frequently occurring amino acids in the N1 and N2 domains of the human repertoire (*e.g.,* G, R, S, P, L, A, V, and T). In other embodiments of the invention, the amino acids used to produce the N1 and N2 segments may be selected from the group consisting of fewer than about 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, or 3 of the amino acids preferentially encoded by the activity of TdT and functionally expressed by human B cells. Alternatively, N1 and N2 may comprise amino acids selected from any group of amino acids. It is not required that N1 and N2 be of a similar length or composition, and independent variation of the length and composition of N1 and N2 is one method by which additional diversity may be introduced into the library.

The DH segments of the libraries are based on the peptides encoded by the naturally occurring IGHD gene repertoire, with progressive deletion of residues at the N- and C-termini. IGHD genes may be read in multiple reading frames, and peptides representing these reading frames, and their N- and C-terminal deletions are also included in the libraries of the invention. In certain embodiments of the invention, DH segments as short as three amino acid residues may be included in the libraries. In other embodiments of the invention, DH segments as short as about 1, 2, 4, 5, 6, 7, or 8 amino acids may be included in the libraries.

The H3-JH segments of the libraries are based on the peptides encoded by the naturally occurring IGHJ gene repertoire, with progressive deletion of residues at the N-terminus. The N-terminal portion of the IGHJ segment that makes up part of the CDRH3 is referred to herein as H3-JH. In certain embodiments of the invention, the H3-JH segment may be represented by progressive N-terminal deletions of one or more H3-JH residues, down to two H3-JH residues. In other embodiments of the invention, the H3-JH segments of the library may contain N-terminal deletions (or no deletions) down to about 6, 5, 4, 3, 2, 1, or 0 H3-JH residues.

The light chain chassis of the libraries may be any sequence with homology to Kabat residues 1 to 88 of naturally occurring light chain (κ or λ) sequences. In certain embodiments of the invention, the light chain chassis of the invention are synthesized in combinatorial fashion, utilizing VL and JL segments, to produce one or more libraries of light chain sequences with diversity in the chassis and CDR3 sequences. In other embodiments of the invention, the light chain CDR3 sequences are synthesized using degenerate oligonucleotides or trinucleotides and recombined with the light chain chassis and light chain constant region, to form full-length light chains.

The instant invention also provides methods for producing and using such libraries, as well as libraries comprising one or more immunoglobulin domains or antibody fragments. Design and synthesis of each component of the claimed antibody libraries is provided in more detail below.

2.1. Design of the Antibody Library Chassis Sequences

One step in building certain libraries of the invention is the selection of chassis sequences, which are based on naturally occurring variable domain sequences (*e.g.,* IGHV and IGLV). This selection can be done arbitrarily, or by the selection of chassis that meet certain criteria. For example, the Kabat database, an electronic database containing non-redundant rearranged antibody sequences, can be queried for those heavy and light chain germline sequences that are most frequently represented. The BLAST search algorithm, or more specialized tools such as SoDA (Volpe *et al.*, Bioinformatics, 2006, 22: 438-44, incorporated by reference in its entirety), can be used to compare rearranged antibody sequences with germline sequences, using the V BASE2 database (Retter *et al.,* Nucleic Acids Res., 2005, 33: D671-D674), or similar collections of human V, D, and J genes, to identify the germline families that are most frequently used to generate functional antibodies.

Several criteria can be utilized for the selection of chassis for inclusion in the libraries of the invention. For example, sequences that are known (or have been determined) to express poorly in yeast, or other organisms used in the invention (*e.g.,* bacteria, mammalian cells, fungi, or plants) can be excluded from the libraries. Chassis may also be chosen based on their representation in the peripheral blood of humans. In certain embodiments of the invention, it may be desirable to select chassis that correspond to germline sequences that are highly represented in the peripheral blood of humans. In other embodiments, it may be desirable to select chassis that correspond to germline sequences that are less frequently represented, for example, to increase the canonical diversity of the library. Therefore, chassis may be selected to produce libraries that represent the largest and most structurally diverse group of functional human antibodies. In other embodiments of the invention, less diverse chassis may be utilized, for example, if it is desirable to produce a smaller, more focused library with less chassis variability and greater CDR variability. In some embodiments of the invention, chassis may be selected based on both their expression in a cell of the invention (*e.g.,* a yeast cell) and the diversity of canonical structures represented by the selected sequences. One may therefore produce a library with a diversity of canonical structures that express well in a cell of the invention.

2.1.1. Design of the Heavy Chain Chassis Sequences

In certain embodiments of the invention, the antibody library comprises variable heavy domains and variable light domains, or portions thereof. Each of these domains is built from certain components, which will be more fully described in the examples provided herein. In certain embodiments, the libraries described herein may be used to isolate fully human antibodies that can be used as diagnostics and/or therapeutics. Without being bound by theory, antibodies with sequences most similar or identical to those most frequently found in peripheral blood (for example, in humans) may be less likely to be immunogenic when administered as therapeutic agents.

Without being bound by theory, and for the purposes of illustrating certain embodiments of the invention, the VH domains of the library may be considered to comprise three primary components: (1) a VH "chassis", which includes amino acids 1 to 94 (using Kabat numbering), (2) the CDRH3, which is defined herein to include the Kabat CDRH3 proper (positions 95-102), and (3) the FRM4 region, including amino acids 103 to 113 (Kabat numbering). The overall VH structure may therefore be depicted schematically (not to scale) as:

```
          (1) … (94)                    (95) … (102)      (103) … (113)
|-------------------------------------|------------------|-------------------------|
          VH Chassis                      CDRH3               FRM4
```

The selection and design of VH chassis sequences based on the human IGHV germline repertoire will become more apparent upon review of the examples provided herein. In certain embodiments of the invention, the VH chassis sequences selected for use in the library may correspond to all functionally expressed human IGHV germline sequences. Alternatively, IGHV germline sequences may be selected for representation in a library according to one or more criteria. For example, in certain embodiments of the invention, the selected IGHV germline sequences may be among those that are most highly represented among antibody molecules isolated from the peripheral blood of healthy adults, children, or fetuses.

In certain embodiments, it may be desirable to base the design of the VH chassis on the utilization of IGHV germline sequences in adults, children, or fetuses with a disease, for example, an autoimmune disease. Without being bound by theory, it is possible that analysis of germline sequence usage in the antibody molecules isolated from the peripheral blood of individuals with autoimmune disease may provide information useful for the design of antibodies recognizing human antigens.

In some embodiments, the selection of IGHV germline sequences for representation in a library of the invention may be based on their frequency of occurrence in the peripheral blood. For the purposes of illustration, four IGHV1 germline sequences (IGHV1-2, IGHV1-18, IGHV1-46, and IGHV1-69) comprise about 80% of the IGHV1 family repertoire in peripheral blood. Thus, the specific IGHV1 germline sequences selected for representation in the library may include those that are most frequently occurring and that cumulatively comprise at least about 80% of the IGHV1 family repertoire found in peripheral blood. An analogous approach can be used to select specific IGHV germline sequences from any other IGHV family (*i.e.,* IGHV1, IGHV2, IGHV3, IGHV4, IGHV5, IGHV6, and IGHV7). The specific germline sequences chosen for representation of a particular IGHV family in a library of the invention may therefore comprise at least about 100%, 99%, 98%, 97%, 96% 95%, 94%, 93%, 92%, 91% 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, 45%, 40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, or 0% of the particular IGHV family member repertoire found in peripheral blood.

In some embodiments, the selected IGHV germline sequences may be chosen to maximize the structural diversity of the VH chassis library. Structural diversity may be evaluated by, for example, comparing the lengths, compositions, and canonical structures of CDRH1 and CDRH2 in the IGHV germline sequences. In human IGHV sequences, the CDRH1 (Kabat definition) may have a length of 5, 6 or 7 amino acids, while CDRH2 (Kabat definition) may have length of 16, 17, 18 or 19 amino acids. The amino acid compositions of the IGHV germline sequences and, in particular, the CDR domains, may be evaluated by sequence alignments, as presented in the Examples. Canonical structure may be assigned, for example, according to the methods described by Chothia *et al.*, J. Mol. Biol., 1992, 227: 799, incorporated by reference in its entirety.

In certain embodiments of the invention, it may be advantageous to design VH chassis based on IGHV germline sequences that may maximize the probability of isolating an antibody with particular characteristics. For example, without being bound by theory, in some embodiments it may be advantageous to restrict the IGHV germline sequences to include only those germline sequences that are utilized in antibodies undergoing clinical development, or antibodies that have been approved as therapeutics. On the other hand, in some embodiments, it may be advantageous to produce libraries containing VH chassis that are not represented amongst clinically utilized antibodies. Such libraries may be capable of yielding antibodies with novel properties that are advantageous over those obtained with the use of "typical" IGHV germline sequences, or enabling studies of the structures and properties of "atypical" IGHV germline sequences or canonical structures.

One of ordinary skill in the art will readily recognize that a variety of other criteria can be used to select IGHV germline sequences for representation in a library of the invention. Any of the criteria described herein may also be combined with any other criteria. Further exemplary criteria include the ability to be expressed at sufficient levels in certain cell culture systems, solubility in particular antibody formats (*e.g.,* whole immunoglobulins and antibody fragments), and the thermodynamic stability of the individual domains, whole immunoglobulins, or antibody fragments. The methods of the invention may be applied to select *any* IGHV germline sequence that has utility in an antibody library of the instant invention.

In certain embodiments of the invention, the VH chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 94 of one or more of the following IGHV germline sequences: IGHV1-2, IGHV1-3, IGHV1-8, IGHV1-18, IGHV1-24, IGHV1-45, IGHV1-46, IGHV1-58, IGHV1-69, IGHV2-5, IGHV2-26, IGHV2-70, IGHV3-7, IGHV3-9, IGHV3-11, IGHV3-13, IGHV3-15, IGHV3-20, IGHV3-21, IGHV3-23, IGHV3-30, IGHV3-33, IGHV3-43, IGHV3-48, IGHV3-49, IGHV3-53, IGHV3-64, IGHV3-66, IGHV3-72, IGHV3-73, IGHV3-74, IGHV4-4, IGHV4-28, IGHV4-31, IGHV4-34, IGHV4-39, IGHV4-59, IGHV4-61, IGHV4-B, IGHV5-51, IGHV6-1, and IGHV7-4-1. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

In other embodiments, the VH chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 94 of the following IGHV germline sequences:

IGHV1-2, IGHV1-18, IGHV1-46, IGHV1-69, IGHV3-7, IGHV3-15, IGHV3-23, IGHV3-30, IGHV3-33, IGHV3-48, IGHV4-31, IGHV4-34, IGHV4-39, IGHV4-59, IGHV4-61, IGHV4-B, and IGHV5-51. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences. The amino acid sequences of these chassis are presented in Table 5.

2.1.1.1. Heavy Chain Chassis Variants

While the selection of the VH chassis with sequences based on the IGHV germline sequences is expected to support a large diversity of CDRH3 sequences, further diversity in the VH chassis may be generated by altering the amino acid residues comprising the CDRH1 and/or CDRH2 regions of each chassis selected for inclusion in the library (see Example 2).

In certain embodiments of the invention, the alterations or mutations in the amino acid residues comprising the CDRH1 and CDRH2 regions, or other regions, of the IGHV germline sequences are made after analyzing the sequence identity within data sets of rearranged human heavy chain sequences that have been classified according to the identity of the original IGHV germline sequence from which the rearranged sequences are derived. For example, from a set of rearranged antibody sequences, the IGHV germline sequence of each antibody is determined, and the rearranged sequences are classified according to the IGHV germline sequence. This determination is made on the basis of sequence identity.

Next, the occurrence of any of the 20 amino acid residues at each position in these sequences is determined. In certain embodiments of the invention, one may be particularly interested in the occurrence of different amino acid residues at the positions within CDRH1 and CDRH2, for example if increasing the diversity of the antigen-binding portion of the VH chassis is desired. In other embodiments of the invention, it may be desirable to evaluate the occurrence of different amino acid residues in the framework regions. Without being bound by theory, alterations in the framework regions may impact antigen binding by altering the spatial orientation of the CDRs.

After the occurrence of amino acids at each position of interest has been identified, alterations may be made in the VH chassis sequence, according to certain criteria. In some embodiments, the objective may be to produce additional VH chassis with sequence variability that mimics the variability observed in the heavy chain domains of rearranged human antibody sequences (derived from respective IGHV germline sequences) as closely as possible, thereby potentially obtaining sequences that are most human in nature (*i.e.,* sequences that most closely mimic the composition and length of human sequences). In this case, one may synthesize additional VH chassis sequences that include mutations naturally found at a particular position and include one or more of these VH chassis sequences in a library of the invention, for example, at a frequency that mimics the frequency found in nature. In another embodiment of the invention, one may wish to include VH chassis that represent only mutations that most frequently occur at a given position in rearranged human antibody sequences. For example, rather than mimicking the human variability precisely, as described above, and with reference to exemplary Tables 6 and 7, one may choose to include only top 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1, amino acid residues that most frequently occur at each position. For the purposes of illustration, and with reference to Table 6, if one wished to include the top four most frequently occurring amino acid residues at position 31 of the VH1-69 sequence, then position 31 in the VH1-69 sequence would be varied to include S, N, T, and R. Without being bound by theory, it is thought that the introduction of diversity by mimicking the naturally occurring composition of the rearranged heavy chain sequences is likely to produce antibodies that are most human in composition. However, the libraries of the invention are not limited to heavy chain sequences that are diversified by this method, and any criteria can be used to introduce diversity into the heavy chain chassis, including random or rational mutagenesis. For example, in certain embodiments of the invention, it may be preferable to substitute neutral and/or smaller amino acid residues for those residues that occur in the IGHV germline sequence. Without being bound by theory, neutral and/or smaller amino acid residues may provide a more flexible and less sterically hindered context for the display of a diversity of CDR sequences.

Example 2 illustrates the application of this method to heavy chains derived from a particular IGHV germline. One of ordinary skill in the art will readily recognize that this method can be applied to any germline sequence, and can be used to generate at least about 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 1000, 10^4, 10^5, 10^6, or more variants of each heavy chain chassis.

2.1.2. Design of the Light Chain Chassis Sequences

The light chain chassis of the invention may be based on kappa and/or lambda light chain sequences. The principles underlying the selection of light chain variable (IGLV) germline sequences for representation in the library are analogous to those employed for the selection of the heavy chain sequences (described above and in Examples 1 and 2). Similarly, the methods used to introduce variability into the selected heavy chain chassis may also be used to introduce variability into the light chain chassis.

Without being bound by theory, and for the purposes of illustrating certain embodiments of the invention, the VL domains of the library may be considered to comprise three primary components: (1) a VL "chassis", which includes amino acids 1 to 88 (using Kabat numbering), (2) the VLCDR3, which is defined herein to include the Kabat CDRL3 proper (positions 89-97), and (3) the FRM4 region, including amino acids 98 to 107 (Kabat numbering). The overall VL structure may therefore be depicted schematically (not to scale) as:

```
        (1) … (88)              (89) … (97)    (98) … (107)
|-------------------------------------|------------------|-------------------------|
        VL Chassis                  CDRL3             FRM4
```

In certain embodiments of the invention, the VL chassis of the libraries include one or more chassis based on IGKV germline sequences. In certain embodiments of the invention, the VL chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 88 of one or more of the following IGKV germline sequences: IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27, IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30, IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-21, and IGKV6D-41. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

In other embodiments, the VL chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 88 of the following IGKV germline sequences: IGKV1-05, IGKV1-12, IGKV1-27, IGKV1-33, IGKV1-39, IGKV2-28, IGKV3-11, IGKV3-15, IGKV3-20, and IGKV4-1. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences. The amino acid sequences of these chassis are presented in Table 11.

In certain embodiments of the invention, the VL chassis of the libraries include one or more chassis based on IGλV germline sequences. In certain embodiments of the invention, the VL chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 88 of one or more of the following IGλV germline sequences: IGλV3-1, IGλV3-21, IGλV2-14, IGλV1-40, IGλV3-19, IGλV1-51, IGλV1-44, IGλV6-57, IGλV2-8, IGλV3-25, IGλV2-23, IGλV3-10, IGλV4-69, IGλV1-47, IGλV2-11, IGλV7-43, IGλV7-46, IGλV5-45, IGλV4-60, IGλV10-54, IGλV8-61, IGλV3-9, IGλV1-36, IGλV2-18, IGλV3-16, IGλV3-27, IGλV4-3, IGλV5-39, IGλV9-49, and IGλV3-12. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

In other embodiments, the VL chassis of the libraries may comprise from about Kabat residue 1 to about Kabat residue 88 of the following IGλV germline sequences: IGλV3-1, IGλV3-21, IGλV2-14, IGλV1-40, IGλV3-19, IGλV1-51, IGλV1-44, IGλV6-57, IGλV4-69, IGλV7-43, and IGλV5-45. In some embodiments of the invention, a library may contain one or more of these sequences, one or more allelic variants of these sequences, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98%, 97%, 96%, 95%, 94%, 93%, 92%, 91%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences. The amino acid sequences of these chassis are presented in Table 14.

2.2. Design of the Antibody Library CDRH3 Components

It is known in the art that diversity in the CDR3 region of the heavy chain is sufficient for most antibody specificities (Xu and Davis, Immunity, 2000, 13: 27-45, incorporated by reference in its entirety) and that existing successful libraries have been created using CDRH3 as the major source of diversification (Hoogenboom *et al.,* J. Mol. Biol., 1992, 227: 381; Lee *et al.*, J. Mol. Biol., 2004, 340: 1073 each of which is incorporated by reference in its entirety). It is also known that both the DH region and the N1/N2 regions contribute to the CDRH3 functional diversity (Schroeder *et al.,* J. Immunol., 2005, 174: 7773 and Mathis *et al.,* Eur J Immunol., 1995, 25: 3115, each of which is incorporated by reference in its entirety). For the purposes of the present invention, the CDHR3 region of naturally occurring human antibodies can be divided into five segments: (1) the tail segment, (2) the N1 segment, (3) the DH segment, (4) the N2 segment, and (5) the JH segment. As exemplified below, the tail, N1 and N2 segments may or may not be present.

In certain embodiments of the invention, the method for selecting amino acid sequences for the synthetic CDRH3 libraries includes a frequency analysis and the generation of the corresponding variability profiles of existing rearranged antibody sequences. In this process, which is described in more detail in the Examples section, the frequency of occurrence of a particular amino acid residue at a particular position within rearranged CDRH3s (or any other heavy or light chain region) is determined. Amino acids that are used more frequently in nature may then be chosen for inclusion in a library of the invention.

2.2.1. Design and Selection of the DH Segment Repertoire

In certain embodiments of the invention, the libraries contain CDRH3 regions comprising one or more segments designed based on the IGHD gene germline repertoire. In some embodiments of the invention, DH segments selected for inclusion in the library are selected and designed based on the most frequent usage of human

IGHD genes, and progressive N-terminal and C-terminal deletions thereof, to mimic the *in vivo* processing of the IGHD gene segments. In some embodiments of the invention, the DH segments of the library are about 3 to about 10 amino acids in length. In some embodiments of the invention, the DH segments of the library are about 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acids in length, or a combination thereof. In certain embodiments, the libraries of the invention may contain DH segments with a wide distribution of lengths (*e.g.,* about 0 to about 10 amino acids). In other embodiments, the length distribution of the DH may be restricted *(e.g.,* about 1 to about 5 amino acids, about 3 amino acids, about 3 and about 5 amino acids, and so on). In certain embodiments of the library, the shortest DH segments may be about 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 amino acids.

In certain embodiments of the invention, libraries may contain DH segments representative of any reading frame of any IGHD germline sequence. In certain embodiments of the invention, the DH segments selected for inclusion in a library include one or more of the following IGHD sequences, or their derivatives (*i.e.,* any reading frame and any degree of N-terminal and C-terminal truncation): IGHD3-10, IGHD3-22, IGHD6-19, IGHD6-13, IGHD3-3, IGHD2-2, IGHD4-17, IGHD1-26, IGHD5-5 / 5-18, IGHD2-15, IGHD6-6, IGHD3-9, IGHD5-12, IGHD5-24, IGHD2-21, IGHD3-16, IGHD4-23, IGHD1-1, IGHD1-7, IGHD4-4/4-11, IGHD1-20, IGHD7-27, IGHD2-8, and IGHD6-25. In some embodiments of the invention, a library may contain one or more of these sequences, allelic variants thereof, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

For the purposes of illustration, progressive N-terminal and C-terminal deletions of IGHD3-10, reading frame 1, are enumerated in the Table 1. N-terminal and C-terminal deletions of other IGHD sequences and reading frames are also encompassed by the invention, and one of ordinary skill in the art can readily determine these sequences using, for example, the non-limiting exemplary data presented in Table 16. and/or the methods outlined above. Table 18 (Example 5) enumerates certain DH segments used in certain embodiments of the invention.

Table 1: Example of Progressive N- and C-terminal Deletions of Reading Frame 1 for Gene IGHD3-10, Yielding DH Segments

DH	SEQ ID NO:	DH	SEQ ID NO:
VLLWFGELL		LWFGEL	
VLLWFGEL		LWFGE	
VLLWFGE		LWFG	
VLLLWFG		LWF	
VLLWF		WFGELL	
VLLW		WFGEL	
VLL		WFGE	
LLWFGELL		WFG	
LLWFGEL		FGELL	
LLWFGE		FGEL	
LLWFG		FGE	
LLWF		GELL	
LLW		GEL	
LWFGELL		ELL	

In certain embodiments of the invention, the DH segments selected for inclusion in a library include one or more of the following IGHD sequences, or their derivatives (*i.e.,* any reading frame and any degree N-terminal and C-terminal truncation): IGHD3-10, IGHD3-22, IGHD6-19, IGHD6-13, IGHD3-03, IGHD2-02, IGHD4-17, IGHD1-26, IGHD5-5/5-18, and IGHD2-15. In some embodiments of the invention, a library may contain one or more of these sequences, allelic variants thereof, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

In certain embodiments of the invention, the DH segments selected for inclusion in a library include one or more of the following IGHD sequences, wherein the notation "_x" denotes the reading frame of the gene, or their derivatives (*i.e.,* any degree of N-terminal or C-terminal truncation): IGHD1-26_1, IGHD1-26_3, IGHD2-2_2, IGHD2-2_3, IGHD2-15_2, IGHD3-3_3, IGHD3-10_1, IGHD3-10_2, IGHD3-10_3, IGHD3-22_2, IGHD4-17_2, IGHD5-5_3, IGHD6-13_1, IGHD6-13_2, IGHD6-19_1, and IGHD6-19_2. In some embodiments of the invention, a library may contain one or more of these sequences, allelic variants thereof, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 55%, or 50% identical to one or more of these sequences.

In certain embodiments of the invention, the libraries are designed to reflect a pre-determined length distribution of N- and C-terminal deleted IGHD segments. For example, in certain embodiments of the library, the DH segments of the library may be designed to mimic the natural length distribution of DH segments found in the human repertoire. For example, the relative occurrence of different IGHD segments in rearranged human antibody heavy chain domains from Lee *et al.* (Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety). Table 2 shows the relative occurrence of the top 68% of IGHD segments from Lee *et al.*

Table 2. Relative Occurrence of Top 68% of IGHD Gene Usage from Lee *et al.*

IGHD Reading Frame	Sequence (Parent)	SEQ ID NO:	Relative Occurrence
IGHD3-10_1	VLLWFGELL		4.3%
IGHD3-10_2	YYYGSGSYYN		8.4%
IGHD3-10_3	ITMVRGVII		4.0%
IGHD3-22_2	YYYDSSGYYY		15.6%
IGHD6-19_1	GYSSGWY		7.4%
IGHD6-19_2	GIAVAG		6.0%
IGHD6-13_1	GYSSSWY		8.4%
IGHD6-13_2	GIAAAG		5.3%
IGHD3-3_3	ITIFGVVII		7.4%
IGHD2-2_2	GYCSSTSCYT		5.2%
IGHD2-2_3	DIVVVPAAM		4.1%
IGHD4-17_2	DYGDY		6.8%
IGHD1-26_1	GIVGATT		2.9%
IGHD1-26_3	YSGSYY		4.3%
IGHD5-5_3	GYSYGY		4.3%
IGHD2-15_2	GYCSGGSCYS		5.6%

In certain embodiments, these relative occurrences may be used to design a library with DH prevalence that is similar to the IGHD usage found in peripheral blood. In other embodiments of the invention, it may be preferable to bias the library toward longer or shorter DH segments, or DH segments of a particular composition. In other embodiments, it may be desirable to use all DH segments selected for the library in equal proportion.

In certain embodiments of the invention, the most commonly used reading-frames of the ten most frequently occurring IGHD sequences are utilized, and progressive N-terminal and C-terminal deletions of these sequences are made, thus providing a total of 278 non-redundant DH segments that are used to create a CDRH3 repertoire of the instant invention (Table 18). In some embodiments of the invention, the methods described above can be applied to produce libraries comprising the top 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 expressed IGHD sequences, and progressive N-terminal and C-terminal deletions thereof. As with all other components of the library, while the DH segments may be selected from among those that are commonly expressed, it is also within the scope of the invention to select these gene segments based on the fact that they are *less* commonly expressed. This may be advantageous, for example, in obtaining antibodies toward self-antigens or in further expanding the diversity of the library. Alternatively, DH segments can be used to add compositional diversity in a manner that is strictly relative to their occurrence in actual human heavy chain sequences.

In certain embodiments of the invention, the progressive deletion of IGHD genes containing disulfide loop encoding segments may be limited, so as to leave the loop intact and to avoid the presence of unpaired cysteine residues. In other embodiments of the invention, the presence of the loop can be ignored and the progressive deletion of the IGHD gene segments can occur as for any other segments, regardless of the presence of unpaired cysteine residues. In still other embodiments of the invention, the cysteine residues can be mutated to any other amino acid.

2.2.2. Design and Selection of the H3-JH Segment Repertoire

There are six IGHJ (joining) segments, IGHJ1, IGHJ2, IGHJ3, IGHJ4, IGHJ5, and IGHJ6. The amino acid sequences of the parent segments and the progressive N-terminal deletions are presented in Table 20 (Example 5). Similar to the N- and C-terminal deletions that the IGHD genes undergo, natural variation is introduced into the IGHJ genes by N-terminal "nibbling", or progressive deletion, of one or more codons by exonuclease activity.

The H3-JH segment refers to the portion of the IGHJ segment that is part of CDRH3. In certain embodiments of the invention, the H3-JH segment of a library comprises one or more of the following sequences: AEYFQH (SEQ ID NO:___), EYFQH (SEQ ID NO: ___), YFQH (SEQ ID NO: ___), FQH (SEQ ID NO: ___), QH (SEQ ID NO: ___), H (SEQ ID NO: ___), YWYFDL (SEQ ID NO: ___), WYFDL (SEQ ID NO: ___), YFDL (SEQ ID NO: ___), FDL (SEQ ID NO: ___), DL (SEQ ID NO: ___), L (SEQ ID NO: ___), AFDV (SEQ ID NO: ___), FDV (SEQ ID NO: ___),

DV (SEQ ID NO: ___), V (SEQ ID NO: ___), YFDY (SEQ ID NO: ___), FDY (SEQ ID NO: ___), DY (SEQ ID NO: ___), Y (SEQ ID NO: ___), NWFDS (SEQ ID NO: ___), WFDS (SEQ ID NO: ___), FDS (SEQ ID NO: ___), DS (SEQ ID NO: ___), S (SEQ ID NO: ___), YYYYYGMDV (SEQ ID NO: ___), YYYYGMDV (SEQ ID NO: ___), YYYGMDV (SEQ ID NO: ___), YYGMDV (SEQ ID NO: ___), YGMDV (SEQ ID NO: ___), GMDV (SEQ ID NO: ___), MDV (SEQ ID NO: ___), and DV (SEQ ID NO: ___). In some embodiments of the invention, a library may contain one or more of these sequences, allelic variations thereof, or encode an amino acid sequence at least about 99.9%, 99.5%, 99%, 98.5%, 98%, 97.5%, 97%, 96.5%, 96%, 95.5%, 95%, 94.5%, 94%, 93.5%, 93%, 92.5%, 92%, 91.5%, 91%, 90.5%, 90%, 89%, 88%, 87%, 86%, 85%, 84%, 83%, 82%, 81%, 80%, 77.5%, 75%, 73.5%, 70%, 65%, 60%, 60%, 55%, or 50% identical to one or more of these sequences.

In other embodiments of the invention, the H3-JH segment may comprise about 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or more amino acids. For example, the H3-JH segment of JH1_4 (Table 20) has a length of three residues, while non-deleted JH6 has an H3-JH segment length of nine residues. The FRM4-JH region of the IGHJ segment begins with the sequence WG(Q/R)G (SEQ ID NO: ___) and corresponds to the portion of the IGHJ segment that makes up part of framework 4. In certain embodiments of the invention, as enumerated in Table 20, there are 28 H3-JH segments that are included in a library. In certain other embodiments, libraries may be produced by utilizing about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 of the IGHJ segments enumerated above or in Table 20.

2.2.3. Design and Selection of the N1 and N2 Segment Repertoires

Terminal deoxynucleotidyl transferase (TdT) is a highly conserved enzyme from vertebrates that catalyzes the attachment of 5' triphosphates to the 3' hydroxyl group of single- or double-stranded DNA. Hence, the enzyme acts as a template-independent polymerase (Koiwai *et al.*, Nucleic Acids Res., 1986, 14: 5777; Basu *et al.,* Biochem. Biophys. Res. Comm., 1983, 111: 1105, each incorporated by reference in its entirety). *In vivo*, TdT is responsible for the addition of nucleotides to the V-D and D-J junctions of antibody heavy chains (Alt and Baltimore, PNAS, 1982, 79: 4118; Collins *et al.,* J. Immunol., 2004, 172: 340, each incorporated by reference in its entirety). Specifically, TdT is responsible for creating the N1 and N2 (non-templated) segments that flank the D (diversity) region.

In certain embodiments of the invention, the length and composition of the N1 and N2 segments are designed rationally, according to statistical biases in amino acid usage found in naturally occurring N1 and N2 segments in human antibodies. One embodiment of a library produced via this method is described in Example 5. According to data compiled from human databases (Jackson *et al.,* J. Immunol Methods, 2007, 324: 26, incorporated by reference in its entirety), there are an average of 3.02 amino acid insertions for N1 and 2.4 amino acid insertions for N2, not taking into account insertions of two nucleotides or less (Figure 2). In certain embodiments of the invention, N1 and N2 segments are restricted to lengths of zero to three amino acids. In other embodiments of the invention, N1 and N2 may be restricted to lengths of less than about 4, 5, 6, 7, 8, 9, or 10 amino acids.

In some embodiments of the invention, the composition of these sequences may be chosen according to the frequency of occurrence of particular amino acids in the N1 and N2 sequences of natural human antibodies (for examples of this analysis, see, Tables 21 to 23, in Example 5). In certain embodiments of the invention, the eight most commonly occurring amino acids in these regions (*i.e.,* G, R, S, P, L, A, T, and V) are used to design the synthetic N1 and N2 segments. In other embodiments of the invention about the most 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, or 19 most commonly occurring amino acids may be used in the design of the synthetic N1 and N2 segments. In still other embodiments, all 20 amino acids may be used in these segments. Finally, while it is possible to base the designed composition of the N1 and N2 segments of the invention on the composition of naturally occurring N1 and N2 segments, this is not a requirement. The N1 and N2 segments may comprise amino acids selected from any group of amino acids, or designed according to other criteria considered for the design of a library of the invention. A person of ordinary skill in the art would readily recognize that the criteria used to design any portion of a library of the invention may vary depending on the application of the particular library. It is an object of the invention that it may be possible to produce a functional library through the use of N1 and N2 segments selected from any group of amino acids, no N1 or N2 segments, or the use of N1 and N2 segments with compositions other than those described herein.

One important difference between the libraries of the current invention and other libraries known in the art is the consideration of the composition of naturally occurring duplet and triplet amino acid sequences during the design of the library. Table 23 shows the top twenty-five naturally occurring duplets in the N1 and N2 regions. Many of these can be represented by the general formula (G/P)(G/R/S/P/L/A/V/T) (SEQ ID NO: ___) or (R/S/L/A/V/T)(G/P) (SEQ ID NO: ___). In certain embodiments of the invention, the synthetic N1 and N2 regions may comprise all of these duplets. In other embodiments, the library may comprise the top 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 most common naturally occurring N1 and/or N2 duplets. In other embodiments of the invention, the libraries may include duplets that are less frequently occurring (*i.e.,* outside of the top 25). The composition of these additional duplets or triplets could readily be determined, given the methods taught herein.

Finally, the data from the naturally occurring triplet N1 and N2 regions demonstrates that the naturally occurring N1 and N2 triplet sequences can often be represented by the formulas (G)(G)(G/R/S/P/L/A/V/T) (SEQ ID NO: ___), (G)(R/S/P/L/A/V/T)(G) (SEQ ID NO: ___), or (R/S/P/L/A/V/T)(G)(G) (SEQ ID NO: ___). In certain embodiments of the invention, the library may comprise the top 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25 most commonly occurring N1 and/or N2 triplets. In other embodiments of the invention, the libraries may include triplets that are less frequently occurring (*i.e.,* outside of the top 25). The composition of these additional duplets or triplets could readily be determined, given the methods taught herein.

In certain embodiments of the invention, there are about 59 total N1 segments and about 59 total N2 segments used to create a library of CDRH3s. In other embodiments of the invention, the number of N1 segments, N2 segments, or both is increased to about 141 (see, for example, Example 5). In other embodiments of the invention, one may select a total of about 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 220, 240, 260, 280, 300, 320, 340, 360, 380, 400, 420, 440, 460, 480, 500, 1000, 10^4, or more N1 and/or N2 segments for inclusion in a library of the invention.

One of ordinary skill in the art will readily recognize that, given the teachings of the instant specification, it is well within the realm of normal experimentation to extend the analysis detailed herein, for example, to generate additional rankings of naturally occurring duplet and triplet (or higher order) N regions that extend beyond those presented herein (*e.g.,* using sequence alignment, the SoDA algorithm, and any database of human sequences (Volpe *et al.*, Bioinformatics, 2006, 22: 438-44, incorporated by reference in its entirety). An ordinarily skilled artisan would also recognize that, based on the information taught herein, it is now possible to produce libraries that are more diverse or less diverse (*i.e.,* more focused) by varying the number of distinct amino acid sequences used in the N1 pool and/or N2 pool.

As described above, many alternative embodiments are envisioned, in which the compositions and lengths of the N1 and N2 segments vary from those presented in the Examples herein. In some embodiments, sub-stoichiometric synthesis of trinucleotides may be used for the synthesis of N1 and N2 segments. Sub-stoichiometric synthesis with trinucleotides is described in Knappik *et al.* (U.S. Patent No. 6,300,064, incorporated by reference in its entirety). The use of sub-stoichiometric synthesis would enable synthesis with consideration of the length variation in the N1 and N2 sequences.

In addition to the embodiments described above, a model of the activity of TdT may also be used to determine the composition of the N1 and N2 sequences in a library of the invention. For example, it has been proposed that the probability of incorporating a particular nucleotide base (A, C, G, T) on a polynucleotide, by the activity of TdT, is dependent on the type of base and the base that occurs on the strand directly preceding the base to be added. Jackson *et al.*, (J. Immunol. Methods, 2007, 324: 26, incorporated by reference in its entirety) have constructed a Markov model describing this process. In certain embodiments of the invention, this model may be used to determine the composition of the N1 and/or N2 segments used in libraries of the invention. Alternatively, the parameters presented in Jackson *et al.* could be further refined to produce sequences that more closely mimic human sequences.

2.2.4. Design of a CDRH3 Library Using the N1, DH, N2, and H3-JH Segments

The CDRH3 libraries of the invention comprise an initial amino acid (in certain exemplary embodiments, G, D, E) or lack thereof (designated herein as position 95), followed by the N1, DH, N2, and H3-JH segments. Thus, in certain embodiments of the invention, the overall design of the CDRH3 libraries can be represented by the following formula:

[G/D/E/-]-[N1]-[DH]-[N2]-[H3-JH].

While the compositions of each portion of a CDRH3 of a library of the invention are more fully described above, the composition of the tail presented above (G/D/E/-) is non-limiting, and that any amino acid (or no amino acid) can be used in this position. Thus, certain embodiments of the invention may be represented by the following formula:

[X]-[N1]-[DH]-[N2]-[H3-JH],

wherein [X] is any amino acid residue or no residue.

In certain embodiments of the invention, a synthetic CDRH3 repertoire is combined with selected VH chassis sequences and heavy chain constant regions, via homologous recombination. Therefore, in certain embodiments of the invention, it may be necessary to include DNA sequences flanking the 5' and 3' ends of the synthetic CDRH3 libraries, to facilitate homologous recombination between the synthetic CDRH3 libraries and vectors containing the selected chassis and constant regions. In certain embodiments, the vectors also contain a sequence encoding at least a portion of the non-nibbled region of the IGHJ gene (*i.e.,* FRM4-JH). Thus, a polynucleotide encoding an N-terminal sequence (*e.g.,* CA(K/R/T)) may be added to the synthetic CDRH3 sequences, wherein the N-terminal polynucleotide is homologous with FRM3 of the chassis, while a polynucleotide encoding a C-terminal sequence (*e.g.,* WG(Q/R)G) may be added to the synthetic CDRH3, wherein the C-terminal polynucleotide is homologous with FRM4-JH. Although the sequence WG(Q/R)G is presented in this exemplary embodiment, additional amino acids, C-terminal to this sequence in FRM4-JH may also be included in the polynucleotide encoding the C-terminal sequence. The purpose of the polynucleotides encoding the N-terminal and C-terminal sequences, in this case, is to facilitate homologous recombination, and one of ordinary skill in the art would recognize that these sequences may be longer or shorter than depicted below. Accordingly, in certain embodiments of the invention, the overall design of the CDRH3 repertoire, including the sequences required to facilitate homologous recombination with the selected chassis, can be represented by the following formula (regions homologous with vector underlined):

CA[R/K/T]-[X]-[N1]-[DH]-[N2]-[H3-JH]-[WG(Q/R)G].

In other embodiments of the invention, the CDRH3 repertoire can be represented by the following formula, which excludes the T residue presented in the schematic above:

CA[R/K]-[X]-[N1]-[DH]-[N2]-[H3-JH]-[WG(Q/R)G].

References describing collections of V, D, and J genes include Scaviner *et al.,* Exp. Clin, Immunogenet., 1999, 16: 243 and Ruiz *et al.,* Exp. Clin. Immunogenet, 1999, 16: 173, each incorporated by reference in its entirety.

2.2.5. CDRH3 Length Distributions

As described throughout this application, in addition to accounting for the composition of naturally occurring CDRH3 segments, the instant invention also takes into account the length distribution of naturally occurring CDRH3 segments. Surveys by Zemlin *et al.* (JMB, 2003, 334: 733, incorporated by reference in its entirety) and Lee *et al.* (Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety) provide analyses of the naturally occurring CDRH3 lengths. These data show that about 95% of naturally occurring CDRH3 sequences have a length from about 7 to about 23 amino acids. In certain embodiments, the instant invention provides rationally designed antibody libraries with CDRH3 segments which directly mimic the size distribution of naturally occurring CDRH3 sequences. In certain embodiments of the invention, the length of the CDRH3s may be about 2 to about 30, about 3 to about 35, about 7 to about 23, about 3 to about 28, about 5 to about 28, about 5 to about 26, about 5 to about 24, about 7 to about 24, about 7 to about 22, about 8 to about 19, about 9 to about 22, about 9 to about 20, about 10 to about 18, about 11 to about 20, about 11 to about 18, about 13 to about 18, or about 13 to about 16 residues in length.

In certain embodiments of the invention, the length distribution of a CDRH3 library of the invention may be defined based on the percentage of sequences within a certain length range. For example, in certain embodiments of the invention, CDRH3s with a length of about 10 to about 18 amino acid residues comprise about 84% to about 94% of the sequences of a the library. In some embodiments, sequences within this length range comprise about 89% of the sequences of a library.

In other embodiments of the invention, CDRH3s with a length of about 11 to about 17 amino acid residues comprise about 74% to about 84% of the sequences of a library. In some embodiments, sequences within this length range comprise about 79% of the sequences of a library.

In still other embodiments of the invention, CDRH3s with a length of about 12 to about 16 residues comprise about 57% to about 67% of the sequences of a library. In some embodiments, sequences within this length range comprise about 62% of the sequences of a library.

In certain embodiments of the invention, CDRH3s with a length of about 13 to about 15 residues comprise about 35% to about 45% of the sequences of a library. In some embodiments, sequences within this length range comprise about 40% of the sequences of a library.

2.3. Design of the Antibody Library CDRL3 Components

The CDRL3 libraries of the invention can be generated by one of several approaches. The actual version of the CDRL3 library made and used in a particular embodiment of the invention will depend on objectives for the use of the library. More than one CDRL3 library may be used in a particular embodiment; for example, a library containing CDRH3 diversity, with kappa and lambda light chains is within the scope of the invention.

In certain embodiments of the invention, a CDRL3 library is a VKCDR3 (kappa) library and/or a VλCDR3 (lambda) library. The CDRL3 libraries described herein differ significantly from CDRL3 libraries in the art. First, they consider length variation that is consistent with what is observed in actual human sequences. Second, they take into consideration the fact that a significant portion of the CDRL3 is encoded by the IGLV gene. Third, the patterns of amino acid variation within the IGLV gene-encoded CDRL3 portions are not stochastic and are selected based on depending on the identity of the IGLV gene. Taken together, the second and third distinctions mean that CDRL3 libraries that faithfully mimic observed patterns in human sequences cannot use a generic design that is independent of the chassis sequences in FRM1 to FRM3. Fourth, the contribution of JL to CDRL3 is also considered explicitly, and enumeration of each amino acid residue at the relevant positions is based on the compositions and natural variations of the JL genes themselves.

As indicated above, and throughout the application, a unique aspect of the design of the libraries of the invention is the germline or "chassis-based" aspect, which is meant to preserve more of the integrity and variability of actual human sequences. This is in contrast to other codon-based synthesis or degenerate oligonucleotide synthesis approaches that have been described in the literature and that aim to produce "one-size-fits-all" (*e.g.,* consensus) libraries (*e.g.,*, Knappik, *et al.*, J Mol Biol, 2000, 296: 57;

Akamatsu *et al.,* J Immunol, 1993, 151: 4651, each incorporated by reference in its entirety).

In certain embodiments of the invention, patterns of occurrence of particular amino acids at defined positions within VL sequences are determined by analyzing data available in public or other databases, for example, the NCBI database (see, for example, GI numbers in Appendices A and B filed herewith). In certain embodiments of the invention, these sequences are compared on the basis of identity and assigned to families on the basis of the germline genes from which they are derived. The amino acid composition at each position of the sequence, in each germline family, may then be determined. This process is illustrated in the Examples provided herein.

2.3.1. Minimalist VKCDR3 Libraries

In certain embodiments of the invention, the light chain CDR3 library is a VKCDR3 library. Certain embodiments of the invention may use only the most common VKCDR3 length, nine residues; this length occurs in a dominant proportion (greater than about 70%) of human VKCDR3 sequences. In human VKCDR3 sequences of length nine, positions 89-95 are encoded by the IGKV gene and positions 96-97 are encoded by the IGKJ gene. Analysis of human kappa light chain sequences indicates that there are not strong biases in the usage of the IGKJ genes. Therefore, in certain embodiments of the invention, each of the five the IGKJ genes can be represented in equal proportions to create a combinatorial library of (M VK chassis) × (5 JK genes), or a library of size $M \times 5$. However, in other embodiments of the invention, it may be desirable to bias IGKJ gene representation, for example to restrict the size of the library or to weight the library toward IGKJ genes known to have particular properties.

As described in Example 6.1, examination of the first amino acid encoded by the IGKJ gene (position 96) indicated that the seven most common residues found at this position are L, Y, R, W, F, P, and I. These residues cumulatively account for about 85% of the residues found in position 96 in naturally occurring kappa light chain sequences. In certain embodiments of the invention, the amino acid residue at position 96 may be one of these seven residues. In other embodiments of the invention, the amino acid at this position may be chosen from amongst any of the other 13 amino acid residues. In still other embodiments of the invention, the amino acid residue at position 96 may be chosen from amongst the top 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 amino acids that occur at position 96, or even residues that never occur at position 96. Similarly, the occurrence of the amino acids selected to occupy position 96 may be equivalent or weighted. In certain embodiments of the invention, it may be desirable to include each of the amino acids selected for inclusion in position 96 at equivalent amounts. In other embodiments of the invention, it may be desirable to bias the composition of position 96 to include particular residues more or less frequently than others. For example, as presented in Example 6.1, arginine occurs at position 96 most frequently when the IGKJ1 germline sequence is used. Therefore, in certain embodiments of the invention, it may be desirable to bias amino acid usage at position 96 according to the origin of the IGKJ germline sequence(s) and/or the IGKV germline sequence(s) selected for representation in a library.

Therefore, in certain embodiments of the invention, a minimalist VKCDR3 library may be represented by one or more of the following amino acid sequences:

[VK_Chassis]-[L3-VK]-[F/L/I/R/W/Y/P]-[JK*]

[VK_Chassis]-[L3-VK]-[X]-[JK*]

In these schematic exemplary sequences, VK_Chassis represents any VK chassis selected for inclusion in a library of the invention (*e.g.,* see Table 11). Specifically, VK_Chassis comprises about Kabat residues 1 to 88 of a selected IGKV sequence. L3-VK represents the portion of the VKCDR3 encoded by the chosen IGKV gene (in this embodiment, Kabat residues 89-95). F, L, I, R, W, Y, and P are the seven most commonly occurring amino acids at position 96 of VKCDR3s with length nine, X is any amino acid, and JK* is an IGKJ amino acid sequence without the N-terminal residue (*i.e.,* the N-terminal residue is substituted with F, L, I, R, W, Y, P, or X). Thus, in one possible embodiment of the minimalist VKCDR3 library, 70 members could be produced by utilizing 10 VK chassis, each paired with its respective L3-VK, 7 amino acids at position 96 (*i.e.,* X), and one JK* sequence. Another embodiment of the library may have 350 members, produced by combining 10 VK chassis, each paired with its respective L3-VK, with 7 amino acids at position 96, and all 5 JK* genes. Still another embodiment of the library may have 1,125 members, produced by combining 15 VK chassis, each paired with its respective H3-JK, with 15 amino acids at position 96 and all

5 JK* genes, and so on. A person of ordinary skill in the art will readily recognize that many other combinations are possible. Moreover, while it is believed that maintaining the pairing between the VK chassis and the L3-VK results in libraries that are more similar to human kappa light chain sequences in composition, the L3-VK regions may also be combinatorially varied with different VK chassis regions, to create additional diversity.

2.3.2. VKCDR3 Libraries of About 10^5 Complexity

While the dominant length of VKCDR3 sequences in humans is about nine amino acids, other lengths appear at measurable frequencies that cumulatively approach almost about 30% of VKCDR3 sequences. In particular, VKCDR3 of lengths 8 and 10 represent about 8.5% and about 16%, respectively, of VKCDR3 lengths in representative samples (Example 6.2; Figure 3). Thus, more complex VKCDR3 libraries may include CDR lengths of 8, 10, and 11 amino acids. Such libraries could account for a greater percentage of the length distribution observed in collections of human VKCDR3 sequences, or even introduce VKCDR3 lengths that do not occur frequently in human VKCDR3 sequences (*e.g.,* less than eight residues or greater than 11 residues).

The inclusion of a diversity of kappa light chain length variations in a library of the invention also enables one to include sequence variability that occurs outside of the amino acid at the VK-JK junction (*i.e.,* position 96, described above). In certain embodiments of the invention, the patterns of sequence variation within the VK, and/or JK segments can be determined by aligning collections of sequences derived from particular germline sequences. In certain embodiments of the invention, the frequency of occurrence of amino acid residues within VKCDR3 can be determined by sequence alignments (*e.g.,* see Example 6.2 and Table 30). In some embodiments of the invention, this frequency of occurrence may be used to introduce variability into the VK_Chassis, L3-VK and/or JK segments that are used to synthesize the VKCDR3 libraries. In certain embodiments of the invention, the top 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 amino acids that occur at any particular position in a naturally occurring repertoire may be included at that position in a VKCDR3 library of the invention. In certain embodiments of the invention, the percent occurrence of any amino acid at any particular position within the VKCDR3 or a VK light chain may be about 0%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100%. In certain embodiments of the invention, the percent occurrence of any amino acid at any position within a VKCDR3 or kappa light chain library of the invention may be within at least about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9% 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 120%, 140%, 160%, 180%, or 200% of the percent occurrence of any amino acid at any position within a naturally occurring VKCDR3 or kappa light chain domain.

In some embodiments of the invention, a VKCDR3 library may be synthesized using degenerate oligonucleotides (see Table 31 for IUPAC base symbol definitions). In some embodiments of the invention, the limits of oligonucleotide synthesis and the genetic code may require the inclusion of more or fewer amino acids at a particular position in the VKCDR3 sequences. An illustrative embodiment of this approach is provided in Example 6.2.

2.3.3. More Complex VKCDR3 Libraries

The limitations inherent in using the genetic code and degenerate oligonucleotide synthesis may, in some cases, require the inclusion of more or fewer amino acids at a particular position within VKCDR3 (*e.g.,* Example 6.2, Table 32), in comparison to those amino acids found at that position in nature. This limitation can be overcome through the use of a codon-based synthesis approach (Virnekas *et al.* Nucleic Acids Res., 1994, 22: 5600, incorporated by reference in its entirety), which enables precise synthesis of oligonucleotides encoding particular amino acids and a finer degree of control over the proportion of any particular amino acid incorporated at any position. Example 6.3 describes this approach in greater detail.

In some embodiments of the invention, a codon-based synthesis approach may be used to vary the percent occurrence of any amino acid at any particular position within the VKCDR3 or kappa light chain. In certain embodiments, the percent occurrence of any amino acid at any position in a VKCDR3 or kappa light chain sequence of the library may be about 0%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, or 100%. In some embodiments of the invention, the percent occurrence of any amino acid at any position may be about 1%, 2%, 3%, or 4%. In certain embodiments of the invention, the percent occurrence of any amino acid at any position within a VKCDR3 or kappa light chain library of the invention may be within at least about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 120%, 140%, 160%, 180%, or 200% of the percent occurrence of any amino acid at any position within a naturally occurring VKCDR3 or kappa light chain domain.

In certain embodiments of the invention, the VKCDR3 (and any other sequence used in the library, regardless of whether or not it is part of VKCDR3) may be altered to remove undesirable amino acid motifs. For example, peptide sequences with the pattern N-X-(S or T)-Z, where X and Z are different from P, will undergo post-translational modification (N-linked glycosylation) in a number of expression systems, including yeast and mammalian cells. In certain embodiments of the invention, the introduction of N residues at certain positions may be avoided, so as to avoid the introduction of N-linked glycosylation sites. In some embodiments of the invention, these modifications may not be necessary, depending on the organism used to express the library and the culture conditions. However, even in the event that the organism used to express libraries with potential N-linked glycosylation sites is incapable of N-linked glycosylation (*e.g.,* bacteria), it may still be desirable to avoid N-X-(S/T) sequences, as the antibodies isolated from such libraries may be expressed in different systems (*e.g.,* yeast, mammalian cells) later (*e.g.,* toward clinical development), and the presence of carbohydrate moieties in the variable domains, and the CDRs in particular, may lead to unwanted modifications of activity.

In certain embodiments of the invention, it may be preferable to create the individual sub-libraries of different lengths (*e.g.,* one or more of lengths 5, 6, 7, 8, 9, 10, 11, or more) separately, and then mix the sub-libraries in proportions that reflect the length distribution of VKCDR3 in human sequences; for example, in ratios approximating the 1:9:2 distribution that occurs in natural VKCDR3 sequences of lengths 8, 9, and 10 (see Figure 3). In other embodiments, it may be desirable to mix these sub-libraries at ratios that are different from the distribution of lengths in natural VKCDR3 sequences, for example, to produce more focused libraries or libraries with particular properties.

2.3.4. VλCDR3 Libraries

The principles used to design the minimalist VλCDR3 libraries of the invention are similar to those enumerated above, for the VKCDR3 libraries, and are explained in more detail in the Examples. One difference between the VλCDR3 libraries of the invention and the VKCDR3 libraries of the invention is that, unlike the IGKV genes, the contribution of the IGVλ genes to CDRL3 (*i.e.,* L3-Vλ) is not constrained to a fixed number of amino acid residues. Therefore, while the combination of the VK (including L3-VK) and JK segments, with inclusion of position 96, yields CDRL3 with a length of only 9 residues, length variation may be obtained within a VλCDR3 library even when only the Vλ (including L3-Vλ) and Jλ segments are considered.

As for the VKCDR3 sequences, additional variability may be introduced into the VλCDR3 sequences via the same methods outlined above, namely determining the frequency of occurrence of particular residues within VλCDR3 sequences and synthesizing the oligonucleotides encoding the desired compositions via degenerate oligonucleotide synthesis or trinucleotides-based synthesis.

2.4. Synthetic Antibody Libraries

In certain embodiments of the invention, both the heavy and light chain chassis sequences and the heavy and light chain CDR3 sequences are synthetic. The polynucleotide sequences of the instant invention can be synthesized by various methods. For example, sequences can be synthesized by split pool DNA synthesis as described in Feldhaus *et al.*, Nucleic Acids Research, 2000, 28: 534; Omstein *et al.,* Biopolymers, 1978, 17: 2341; and Brenner and Lerner, PNAS, 1992, 87: 6378 (each of which is incorporated by reference in its entirety).

In some embodiments of the invention, cassettes representing the possible V, D, and J diversity found in the human repertoire, as well as junctional diversity, are synthesized *de novo* either as double-stranded DNA oligonucleotides, single-stranded DNA oligonucleotides representative of the coding strand, or single-stranded DNA oligonucleotides representative of the non-coding strand. These sequences can then be introduced into a host cell along with an acceptor vector containing a chassis sequence and, in some cases a portion of FRM4 and a constant region. No primer-based PCR amplification from mammalian cDNA or mRNA or template-directed cloning steps from mammalian cDNA or mRNA need be employed.

2.5. Construction of Libraries by Yeast Homologous Recombination

In certain embodiments, the present invention exploits the inherent ability of yeast cells to facilitate homologous recombination at high efficiency. The mechanism of homologous recombination in yeast and its applications are briefly described below.

As an illustrative embodiment, homologous recombination can be carried out in, for example, *Saccharomyces cerevisiae*, which has genetic machinery designed to carry out homologous recombination with high efficiency. Exemplary *S. cerevisiae* strains include EM93, CEN.PK2, RM11-1a, YJM789, and BJ5465. This mechanism is believed to have evolved for the purpose of chromosomal repair, and is also called "gap repair" or "gap filling". By exploiting this mechanism, mutations can be introduced into specific loci of the yeast genome. For example, a vector carrying a mutant gene can contain two sequence segments that are homologous to the 5' and 3' open reading frame (ORF) sequences of a gene that is intended to be interrupted or mutated. The vector may also encode a positive selection marker, such as a nutritional enzyme allele (*e.g.,* URA3) and/or an antibiotic resistant marker (*e.g.,* Geneticin / G418), flanked by the two homologous DNA segments. Other selection markers and antibiotic resistance markers are known to one of ordinary skill in the art. In some embodiments of the invention, this vector (*e.g.,* a plasmid) is linearized and transformed into the yeast cells. Through homologous recombination between the plasmid and the yeast genome, at the two homologous recombination sites, a reciprocal exchange of the DNA content occurs between the wild type gene in the yeast genome and the mutant gene (including the selection marker gene(s)) that is flanked by the two homologous sequence segments. By selecting for the one or more selection markers, the surviving yeast cells will be those cells in which the wild-type gene has been replaced by the mutant gene (Pearson *et al.,* Yeast, 1998, 14: 391, incorporated by reference in its entirety). This mechanism has been used to make systematic mutations in all 6,000 yeast genes, or open reading frames (ORFs), for functional genomics studies. Because the exchange is reciprocal, a similar approach has also been used successfully to clone yeast genomic DNA fragments into a plasmid vector (Iwasaki *et al.,* Gene, 1991, 109: 81, incorporated by reference in its entirety).

By utilizing the endogenous homologous recombination machinery present in yeast, gene fragments or synthetic oligonucleotides can also be cloned into a plasmid vector without a ligation step. In this application of homologous recombination, a target gene fragment (*i.e.,* the fragment to be inserted into a plasmid vector, *e.g.,* a CDR3) is obtained (*e.g.,* by oligonucleotides synthesis, PCR amplification, restriction digestion out of another vector, etc.). DNA sequences that are homologous to selected regions of the plasmid vector are added to the 5' and 3' ends of the target gene fragment. These homologous regions may be fully synthetic, or added via PCR amplification of a target gene fragment with primers that incorporate the homologous sequences. The plasmid vector may include a positive selection marker, such as a nutritional enzyme allele (*e.g.,* URA3), or an antibiotic resistance marker (*e.g.,* Geneticin / G418). The plasmid vector is then linearized by a unique restriction cut located in-between the regions of sequence homology shared with the target gene fragment, thereby creating an artificial gap at the cleavage site. The linearized plasmid vector and the target gene fragment flanked by sequences homologous to the plasmid vector are co-transformed into a yeast host strain. The yeast is then able to recognize the two stretches of sequence homology between the vector and target gene fragment and facilitate a reciprocal exchange of DNA content through homologous recombination at the gap. As a consequence, the target gene fragment is inserted into the vector without ligation.

The method described above has also been demonstrated to work when the target gene fragments are in the form of single stranded DNA, for example, as a circular M13 phage derived form, or as single stranded oligonucleotides (Simon and Moore, Mol. Cell Biol., 1987, 7: 2329; Ivanov *et al.,* Genetics, 1996, 142: 693; and DeMarini *et al.,* 2001, 30: 520., each incorporated by reference in its entirety). Thus, the form of the target that can be recombined into the gapped vector can be double stranded or single stranded, and derived from chemical synthesis, PCR, restriction digestion, or other methods.

Several factors may influence the efficiency of homologous recombination in yeast. For example, the efficiency of the gap repair is correlated with the length of the homologous sequences flanking both the linearized vector and the target gene. In certain embodiments, about 20 or more base pairs may be used for the length of the homologous sequence, and about 80 base pairs may give a near-optimized result (Hua *et al.,* Plasmid, 1997, 38: 91; Raymond *et al.,* Genome Res., 2002, 12: 190, each incorporated by reference in its entirety). In certain embodiments of the invention, at least about 5, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 35, 36, 37, 38, 39, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 187, 190, or 200 homologous base pairs may be used to facilitate recombination. In other embodiments, between about 20 and about 40 base pairs are utilized. In addition, the reciprocal exchange between the vector and gene fragment is strictly sequence-dependent, *i.e.* it does not cause a frame shift. Therefore, gap-repair cloning assures the insertion of gene fragments with both high efficiency and precision. The high efficiency makes it possible to clone two, three, or more targeted gene fragments simultaneously into the same vector in one transformation attempt (Raymond *et al.,* Biotechniques, 1999, 26: 134, incorporated by reference in its entirety). Moreover, the nature of precision sequence conservation through homologous recombination makes it possible to clone selected genes or gene fragments into expression or fusion vectors for direct functional examination (El-Deiry *et al.,* Nature Genetics, 1992, 1: 4549; Ishioka *et al.,* PNAS, 1997, 94: 2449, each incorporated by reference in its entirety).

Libraries of gene fragments have also been constructed in yeast using homologous recombination. For example, a human brain cDNA library was constructed as a two-hybrid fusion library in vector pJG4-5 (Guidotti and Zervos, Yeast, 1999, 15: 715, incorporated by reference in its entirety). It has also been reported that a total of 6,000 pairs of PCR primers were used for amplification of 6,000 known yeast ORFs for a study of yeast genomic protein interactions (Hudson *et al.,* Genome Res., 1997, 7: 1169, incorporated by reference in its entirety). In 2000, Uetz *et al.* conducted a comprehensive analysis-of protein-protein interactions in *Saccharomyces cerevisiae* (Uetz *et al.,* Nature, 2000, 403: 623, incorporated by reference in its entirety). The protein-protein interaction map of the budding yeast was studied by using a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins (Ito *et al.,* PNAS, 2000, 97: 1143, incorporated by reference in its entirety), and the genomic protein linkage map of Vaccinia virus was studied using this system (McCraith *et al.,* PNAS, 2000, 97: 4879, incorporated by reference in its entirety).

In certain embodiments of the invention, a synthetic CDR3 (heavy or light chain) may be joined by homologous recombination with a vector encoding a heavy or light chain chassis, a portion of FRM4, and a constant region, to form a full-length heavy or light chain. In certain embodiments of the invention, the homologous recombination is performed directly in yeast cells. In some embodiments, the method comprises:

(a) transforming into yeast cells:

(i) a linearized vector encoding a heavy or light chain chassis, a portion of FRM4, and a constant region, wherein the site of linearization is between the end of FRM3 of the chassis and the beginning of the constant region; and

(ii) a library of CDR3 insert nucleotide sequences that are linear and double stranded, wherein each of the CDR3 insert sequences comprises a nucleotide sequence encoding CDR3 and 5'- and 3'-flanking sequences that are sufficiently homologous to the termini of the vector of (i) at the site of linearization to enable homologous recombination to occur between the vector and the library of CDR3 insert sequences; and

(b) allowing homologous recombination to occur between the vector and the CDR3 insert sequences in the transformed yeast cells, such that the CDR3 insert sequences are incorporated into the vector, to produce a vector encoding full-length heavy chain or light chain.

As specified above, the CDR3 inserts may have a 5' flanking sequence and a 3' flanking sequence that are homologous to the termini of the linearized vector. When the CDR3 inserts and the linearized vectors are introduced into a host cell, for example, a yeast cell, the "gap" (the linearization site) created by linearization of the vector is filled by the CDR3 fragment insert through recombination of the homologous sequences at the 5' and 3' termini of these two linear double-stranded DNAs (*i.e.,* the vector and the insert). Through this event of homologous recombination, libraries of circular vectors encoding full-length heavy or light chains comprising variable CDR3 inserts is generated. Particular instances of these methods are presented in the Examples.

Subsequent analysis may be carried out to determine the efficiency of homologous recombination that results in correct insertion of the CDR3 sequences into the vectors. For example, PCR amplification of the CDR3 inserts directly from selected yeast clones may reveal how many clones are recombinant. In certain embodiments, libraries with minimum of about 90% recombinant clones are utilized. In certain other embodiments libraries with a minimum of about 1%, 5% 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% recombinant clones are utilized. The same PCR amplification of selected clones may also reveal the insert size.

To verify the sequence diversity of the inserts in the selected clones, a PCR amplification product with the correct size of insert may be "fingerprinted" with restriction enzymes known to cut or not cut within the amplified region. From a gel electrophoresis pattern, it may be determined whether the clones analyzed are of the same identity or of the distinct or diversified identity. The PCR products may also be sequenced directly to reveal the identity of inserts and the fidelity of the cloning procedure, and to prove the independence and diversity of the clones. Figure 1 depicts a schematic of recombination between a fragment (*e.g.,* CDR3) and a vector (*e.g.,* comprising a chassis, portion of FRM4, and constant region) for the construction of a library.

2.6. Expression and Screening Systems

Libraries of polynucleotides generated by any of the techniques described herein, 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, for example, using cell-free extracts (and *e.g.*, ribosome display), phage display, prokaryotic cells (*e.g.,* bacterial display), or eukaryotic cells (*e.g.*, yeast display). In certain embodiments of the invention, the antibody libraries are expressed in yeast.

In other embodiments, 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, (each incorporated by reference in its entirety) 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, each incorporated by reference in its entirety).

Alternatively, the polynucleotides of the invention can be expressed in an *E. coli* expression system, such as that described by Pluckthun and Skerra. (Meth. Enzymol., 1989, 178: 476; Biotechnology, 1991, 9: 273, each incorporated by reference in its entirety). The mutant proteins can be expressed for secretion in the medium and/or in the cytoplasm of the bacteria, as described by Better and Horwitz, Meth. Enzymol., 1989, 178: 476, incorporated by reference in its entirety. In some embodiments, 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 *et al.*, J. Bacteriol., 1987, 169: 4379, incorporated by reference in its entirety). 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 *et al.*, Biotechnology, 1991, 9: 273, incorporated by reference in its entirety). For example, antibody heavy chain genes can be concurrently expressed with antibody light chain genes to produce antibodies or antibody fragments.

In other embodiments of the invention, 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 US20040072740; US20030100023; and US20030036092 (each incorporated by reference in its entirety).

Higher eukaryotic cells, such as mammalian cells, for example myeloma cells (*e.g.*, NS/0 cells), hybridoma cells, Chinese hamster ovary (CHO), and human embryonic kidney (HEK) cells, can also be used for expression of the antibodies of the invention. Typically, antibodies expressed in mammalian cells are designed to be secreted into the culture medium, or expressed on the surface of the cell. The antibody or antibody fragments can be produced, for example, as intact antibody molecules or as individual VH and VL fragments, Fab fragments, single domains, or as single chains (scFv) (Huston *et al.*, PNAS, 1988, 85: 5879, incorporated by reference in its entirety).

Alternatively, antibodies can be expressed and screened by anchored periplasmic expression (APEx 2-hybrid surface display), as described, for example, in Jeong *et al.*, PNAS, 2007, 104: 8247 (incorporated by reference in its entirety) or by other anchoring methods as described, for example, in Mazor *et al.*, Nature Biotechnology, 2007, 25: 563 (incorporated by reference in its entirety).

In other embodiments of the invention, antibodies can be selected using mammalian cell display (Ho *et al.*, PNAS, 2006, 103: 9637, incorporated by reference in its entirety).

The screening of the antibodies derived from the libraries of the invention can be carried out 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 assays, *e.g.*, the hemoglobin plaque assay as described in U.S. Patent No.

5,798,208 (incorporated by reference in its entirety). Determining the ability of candidate antibodies to bind therapeutic targets can be assayed *in vitro* using, *e.g.*, a BIACORE^{TM} instrument, which measures binding rates of an antibody to a given target or antigen based on surface plasmon resonance. *In vivo* assays can be conducted using any of a number of animal models and then subsequently tested, as appropriate, in humans. Cell-based biological assays are also contemplated.

One aspect of the instant invention is the speed at which the antibodies of the library can be expressed and screened. In certain embodiments of the invention, the antibody library can be expressed in yeast, which have a doubling time of less than about 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, or 24 hours. In some embodiments, the doubling times are about 1 to about 3 hours, about 2 to about 4, about 3 to about 8 hours, about 3 to about 24, about 5 to about 24, about 4 to about 6 about 5 to about 22, about 6 to about 8, about 7 to about 22, about 8 to about 10 hours, about 7 to about 20, about 9 to about 20, about 9 to about 18, about 11 to about 18, about 11 to about 16, about 13 to about 16, about 16 to about 20, or about 20 to about 30 hours. In certain embodiments of the invention, the antibody library is expressed in yeast with a doubling time of about 16 to about 20 hours, about 8 to about 16 hours, or about 4 to about 8 hours. Thus, the antibody library of the instant invention can be expressed and screened in a matter of hours, as compared to previously known techniques which take several days to express and screen antibody libraries. A limiting step in the throughput of such screening processes in mammalian cells is simply the time required to iteratively regrow populations of isolated cells, which, in some cases, have doubling times greater than the doubling times of the yeast used in the current invention.

In certain embodiments of the invention, the composition of a library may be defined after one or more enrichment steps (for example by screening for antigen binding, or other properties). For example, a library with a composition comprising about x% sequences or libraries of the invention may be enriched to contain about 2x%, 3x%, 4x%, 5x%, 6x%, 7x%, 8x%, 9x%, 10x%, 20x%, 25x%, 40x%, 50x%, 60x% 75x%, 80x%, 90x%, 95x%,or 99x% sequences or libraries of the invention, after one or more screening steps. In other embodiments of the invention, the sequences or libraries of the invention may be enriched about 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, 100-fold, 1,000-fold, or more, relative to their occurrence prior to the one or more enrichment steps. In certain embodiments of the invention, a library may contain at least a certain number of a particular type of sequence(s), such as CDRH3s, CDRL3s, heavy chains, light chains, or whole antibodies (*e.g.,* at least about 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, or 10^{20}). In certain embodiments, these sequences may be enriched during one or more enrichment steps, to provide libraries comprising at least about 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, or 10^{19} of the respective sequence(s).

2.7. Mutagenesis Approaches for Affinity Maturation

As described above, antibody leads can be identified through a selection process that involves screening the antibodies of a library of the invention for binding to one or more antigens, or for a biological activity. The coding sequences of these antibody leads may be further mutagenized *in vitro* or *in vivo* to generate secondary libraries with diversity introduced in the context of the initial antibody leads. The mutagenized antibody leads can then be further screened for binding to target antigens or biological activity, *in vitro* or *in vivo*, following procedures similar to those used for the selection of the initial antibody lead from the primary library. Such mutagenesis and selection of primary antibody leads effectively mimics the affinity maturation process naturally occurring in a mammal that produces antibodies with progressive increases in the affinity to an antigen. In one embodiment of the invention, only the CDRH3 region is mutagenized. In another embodiment of the invention, the whole variable region is mutagenized. In other embodiments of the invention one or more of CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, and/ CDRL3 may be mutagenized. In some embodiments of the invention, "light chain shuffling" may be used as part of the affinity maturation protocol. In certain embodiments, this may involve pairing one or more heavy chains with a number of light chains, to select light chains that enhance the affinity and/or biological activity of an antibody. In certain embodiments of the invention, the number of light chains to which the one or more heavy chains can be paired is at least about 2, 5, 10, 100, 1000, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, or 10^{10}. In certain embodiments of the invention, these light chains are encoded by plasmids. In other embodiments of the invention, the light chains may be integrated into the genome of the host cell.

The coding sequences of the antibody leads may be mutagenized by a wide variety of methods. Examples of methods of mutagenesis include, but are not limited to site-directed mutagenesis, error-prone PCR mutagenesis, cassette mutagenesis, and random PCR mutagenesis. Alternatively, oligonucleotides encoding regions with the desired mutations can be synthesized and introduced into the sequence to be mutagenized, for example, via recombination or ligation.

Site-directed mutagenesis or point mutagenesis may be used to gradually change the CDR sequences in specific regions. This may be accomplished by using oligonucleotide-directed mutagenesis or PCR. For example, a short sequence of an antibody lead may be replaced with a synthetically mutagenized oligonucleotide in either the heavy chain or light chain region, or both. The method may not be efficient for mutagenizing large numbers of CDR sequences, but may be used for fine tuning of a particular lead to achieve higher affinity toward a specific target protein.

Cassette mutagenesis may also be used to mutagenize the CDR sequences in specific regions. In a typical cassette mutagenesis, a sequence block, or a region, of a single template is replaced by a completely or partially randomized sequence. However, the maximum information content that can be obtained may be statistically limited by the number of random sequences of the oligonucleotides. Similar to point mutagenesis, this method may also be used for fine tuning of a particular lead to achieve higher affinity towards a specific target protein.

Error-prone PCR, or "poison" PCR, may be used to mutagenize the CDR sequences by following protocols described in Caldwell and Joyce, PCR Methods and Applications, 1992, 2: 28; Leung *et al.,* Technique, 1989, 1: 11; Shafikhani *et al.,* Biotechniques, 1997, 23: 304; and Stemmer *et al.,* PNAS, 1994, 91: 10747 (each of which is incorporated by reference in its entirety).

Conditions for error prone PCR may include (a) high concentrations of Mn^{2+} (*e.g.,* about 0.4 to about 0.6 mM) that efficiently induces malfunction of Taq DNA polymerase; and (b) a disproportionally high concentration of one nucleotide substrate (*e.g.,* dGTP) in the PCR reaction that causes incorrect incorporation of this high concentration substrate into the template and produces mutations. Additionally, other factors such as, the number of PCR cycles, the species of DNA polymerase used, and the length of the template, may affect the rate of misincorporation of "wrong" nucleotides into the PCR product. Commercially available kits may be utilized for the mutagenesis of the selected antibody library, such as the "Diversity PCR random mutagenesis kit" (CLONTECH™).

The primer pairs used in PCR-based mutagenesis may, in certain embodiments, include regions matched with the homologous recombination sites in the expression vectors. This design allows facile re-introduction of the PCR products back into the heavy or light chain chassis vectors, after mutagenesis, via homologous recombination.

Other PCR-based mutagenesis methods can also be used, alone or in conjunction with the error prone PCR described above. For example, the PCR amplified CDR segments may be digested with DNase to create nicks in the double stranded DNA. These nicks can be expanded into gaps by other exonucleases such as Bal 31. The gaps may then be filled by random sequences by using DNA Klenow polymerase at a low concentration of regular substrates dGTP, dATP, dTTP, and dCTP with one substrate (e.g., dGTP) at a disproportionately high concentration. This fill-in reaction should produce high frequency mutations in the filled gap regions. These method of DNase digestion may be used in conjunction with error prone PCR to create a high frequency of mutations in the desired CDR segments.

The CDR or antibody segments amplified from the primary antibody leads may also be mutagenized *in vivo* by exploiting the inherent ability of mutation in pre-B cells. The Ig genes in pre-B cells are specifically susceptible to a high-rate of mutation. The Ig promoter and enhancer facilitate such high rate mutations in a pre-B cell environment while the pre-B cells proliferate. Accordingly, CDR gene segments may be cloned into a mammalian expression vector that contains a human Ig enhancer and promoter. This construct may be introduced into a pre-B cell line, such as 38B9, which allows the mutation of the VH and VL gene segments naturally in the pre-B cells (Liu and Van Ness, Mol. Immunol., 1999, 36: 461, incorporated by reference in its entirety). The mutagenized CDR segments can be amplified from the cultured pre-B cell line and re-introduced back into the chassis-containing vector(s) via, for example, homologous recombination.

In some embodiments, a CDR “hit” isolated from screening the library can be re-synthesized, using degenerate codons or trinucleotides, and re-cloned into the heavy or light chain vector using gap repair.

3. Library Sampling

In certain embodiments of the invention, a library of the invention comprises a designed, non-random repertoire wherein the theoretical diversity of particular components of the library (for example, CDRH3), but not necessarily all components or the entire library, can be over-sampled in a physical realization of the library, at a level where there is a certain degree of statistical confidence (*e.g.,* 95%) that any given member of the theoretical library is present in the physical realization of the library at least at a certain frequency (*e.g.,* at least once, twice, three times, four times, five times, or more) in the library.

In a library, it is generally assumed that the number of copies of a given clone obeys a Poisson probability distribution (see Feller, W. *An Introduction to Probability Theory and Its Applications*, 1968, Wiley New York, incorporated by reference in its entirety). The probability of a Poisson random number being zero, corresponding to the probability of missing a given component member in an instance of a library (see below), is e^{-N}, where N is the average of the random number. For example, if there are 10^6 possible theoretical members of a library and a physical realization of the library has 10^7 members, with an equal probability of each member of the theoretical library being sampled, then the average number of times that each member occurs in the physical realization of the library is $10^7/10^6=10$, and the probability that the number of copies of a given member is zero is $e^{-N}=e^{-10}=0.000045$; or a 99.9955% chance that there is at least one copy of any of the 10^6 theoretical members in this 10X oversampled library. For a 2.3X oversampled library one is 90% confident that a given component is present. For a 3X oversampled library one is 95% confident that a given component is present. For a 4.6X oversampled library one is 99% confident a given clone is present, and so on.

Therefore, if M is the maximum number of theoretical library members that can be feasibly physically realized, then $M/3$ is the maximum theoretical repertoire size for which one can be 95% confident that any given member of the theoretical library will be sampled. It is important to note that there is a difference between a 95% chance that a given member is represented and a 95% chance that every possible member is represented. In certain embodiments, the instant invention provides a rationally designed library with diversity so that any given member is 95% likely to be represented in a physical realization of the library. In other embodiments of the invention, the library is designed so that any given member is at least about 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 85%, 90%, 95%, 99%, 99.5%, or 99.9% likely to be represented in a physical realization of the library. For a review, see, *e.g.,* Firth and Patrick, Biomol. Eng., 2005, 22: 105, and Patrick *et al.,* Protein Engineering, 2003, 16: 451, each of which is incorporated by reference in its entirety.

In certain embodiments of the invention, a library may have a theoretical total diversity of X unique members and the physical realization of the theoretical total diversity may contain at least about 1X, 2X, 3X, 4X, 5X, 6X, 7X, 8X 9X, 10X, or more members. In some embodiments, the physical realization of the theoretical total diversity may contain about 1X to about 2X, about 2X to about 3X, about 3X to about 4X, about 4X to about 5X, about 5X to about 6X members. In other embodiments, the physical realization of the theoretical total diversity may contain about 1X to about 3X, or about 3X to about 5X total members.

An assumption underlying all directed evolution experiments is that the amount of molecular diversity theoretically possible is enormous compared with the ability to synthesize it, physically realize it, and screen it. The likelihood of finding a variant with improved properties in a given library is maximized when that library is maximally diverse. Patrick et al. used simple statistics to derive a series of equations and computer algorithms for estimating the number of unique sequence variants in libraries constructed by randomized oligonucleotide mutagenesis, error-prone PCR and in vitro recombination. They have written a suite of programs for calculating library statistics, such as GLUE, GLUE-IT, PEDEL, PEDEL-AA, and DRIVeR. These programs are described, with instructions on how to access them, in Patrick *et al.,* Protein Engineering, 2003, 16: 451and Firth *et al.,* Nucleic Acids Res., 2008, 36: W281 (each of which is incorporated by reference in its entirety).

It is possible to construct a physical realization of a library in which some components of the theoretical diversity (such as CDRH3) are oversampled, while other aspects (VH/VL pairings) are not. For example, consider a library in which 10^8 CDRH3 segments are designed to be present in a single VH chassis, and then paired with 10^5 VL genes to produce 10^{13} (= $10^8 * 10^5$) possible full heterodimeric antibodies. If a physical realization of this library is constructed with a diversity of 10^9 transformant clones, then the CDRH3 diversity is oversampled ten-fold (= $10^9/10^8$), however the possible VH/VL pairings are undersampled by 10^{-4} (= $10^9/10^{13}$). In this example, on average, each CDRH3 is paired only with 10 samples of the VL from the possible 10^5 partners. In certain embodiments of the invention, it is the CDRH3 diversity that is preferably oversampled.

3.1. Other Variants of the Polynucleotide Sequences of the Invention

In certain embodiments, the invention relates to a polynucleotide that hybridizes with a polynucleotide taught herein, or that hybridizes with the complement of a polynucleotide taught herein. For example, an isolated polynucleotide that remains hybridized after hybridization and washing under low, medium, or high stringency conditions to a polynucleotide taught herein or the complement of a polynucleotide taught herein is encompassed by the present invention.

Exemplary low stringency conditions include hybridization with a buffer solution of about 30% to about 35% formamide, about 1 M NaCl, about 1% SDS (sodium dodecyl sulphate) at about 37°C, and a wash in about 1X to about 2X SSC (20X SSC=3.0 M NaCl/0.3 M trisodium citrate) at about 50°C to about 55°C.

Exemplary moderate stringency conditions include hybridization in about 40% to about 45% formamide, about 1 M NaCl, about 1% SDS at about 37°C, and a wash in about 0.5X to about 1X SSC at abut 55°C to about 60°C.

Exemplary high stringency conditions include hybridization in about 50% formamide, about 1 M NaCl, about 1% SDS at about 37°C, and a wash in about 0.1X SSC at about 60° C to about 65° C.

Optionally, wash buffers may comprise about 0.1% to about 1% SDS.

The duration of hybridization is generally less than about 24 hours, usually about 4 to about 12 hours.

3.2. Sub-Libraries and Larger Libraries Comprising the Libraries or Sub-Libraries of the Invention

As described throughout the application, the libraries of the current invention are distinguished, in certain embodiments, by their human-like sequence composition and length, and the ability to generate a physical realization of the library which contains all members of (or, in some cases, even oversamples) a particular component of the library. Libraries comprising combinations of the libraries described herein (*e.g.,* CDRH3 and CDRL3 libraries) are encompassed by the invention. Sub-libraries comprising portions of the libraries described herein are also encompassed by the invention (*e.g.,* a CDRH3 library in a particular heavy chain chassis or a sub-set of the CDRH3 libraries). One of ordinary skill in the art will readily recognize that each of the libraries described herein has several components (*e.g.,* CDRH3, VH, CDRL3, VL, etc.), and that the diversity of these components can be varied to produce sub-libraries that fall within the scope of the invention.

Moreover, libraries containing one of the libraries or sub-libraries of the invention also fall within the scope of the invention. For example, in certain embodiments of the invention, one or more libraries or sub-libraries of the invention may be contained within a larger library, which may include sequences derived by other means, for example, non-human or human sequence derived by stochastic or semi-stochastic synthesis. In certain embodiments of the invention, at least about 1% of the sequences in a polynucleotide library may be those of the invention (*e.g.,* CDRH3 sequences, CDRL3 sequences, VH sequences, VL sequences), regardless of the composition of the other 99% of sequences. In other embodiments of the invention, at least about 0.001%, 0.01%, 0.1%, 2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91,%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% of the sequences in any polynucleotide library may be those of the invention, regardless of the composition of the other sequences. In some embodiments, the sequences of the invention may comprise about 0.001% to about 1%, about 1% to about 2%, about 2% to about 5%, about 5% to about 10%, about 10% to about 15%, about 15% to about 20%, about 20% to about 25%, about 25% to about 30%, about 30% to about 35%, about 35% to about 40%, about 40% to about 45%, about 45% to about 50%, about 50% to about 55%, about 55% to about 60%, about 60% to about 65%, about 65% to about 70%, about 70% to about 75%, about 75% to about 80%, about 80% to about 85%, about 85% to about 90%, about 90% to about 95%, or about 95% to about 99% of the sequences in any polynucleotide library, regardless of the composition of the other sequences. Thus, libraries more diverse than one or more libraries or sub-libraries of the invention, but yet still comprising one or more libraries or sub-libraries of the invention, in an amount in which the one or more libraries or sub-libraries of the invention can be effectively screened and from which sequences encoded by the one or more libraries or sub-libraries of the invention can be isolated, also fall within the scope of the invention.

3.3. Alternative Scaffolds

In certain embodiments of the invention, the amino acid products of a library of the invention (*e.g.,* a CDRH3 or CDRL3) may be displayed on an alternative scaffold. Several of these scaffolds have been shown to yield molecules with specificities and affinities that rival those of antibodies. Exemplary alternative scaffolds include those derived from fibronectin (*e.g.,* AdNectin), the β-sandwich (*e.g.,* iMab), lipocalin (*e.g.,* Anticalin), EETI-II/AGRP, BPTI/LACI-D1/ITI-D2 (*e.g.,* Kunitz domain), thioredoxin (*e.g.,* peptide aptamer), protein A (*e.g.,* Affibody), ankyrin repeats (*e.g.,* DARPin), γB-crystallin/ubiquitin (*e.g.,* Affilin), CTLD$_3$ (*e.g.,* Tetranectin), and (LDLR-A module)$_3$ (*e.g.,* Avimers). Additional information on alternative scaffolds are provided in Binz *et al.,* Nat. Biotechnol., 2005 23: 1257 and Skerra, Current Opin. in Biotech., 2007 18: 295-304, each of which is incorporated by reference in its entirety.

4. Other Embodiments of the Invention

In certain embodiments, the invention comprises a synthetic preimmune human antibody CDRH3 library comprising 10^7 to 10^8 polynucleotide sequences representative of the sequence diversity and length diversity found in known heavy chain CDR3 sequences.

In other embodiments, the invention comprises a synthetic preimmune human antibody CDRH3 library comprising polynucleotide sequences encoding CDRH3 represented by the following formula:

[G/D/E/-][N1][DH][N2][H3-JH],

wherein [G/D/E/-] is zero to one amino acids in length, [N1] is zero to three amino acids, [DH] is three to ten amino acids in length, [N2] is zero to three amino acids in length, and [H3-JH] is two to nine amino acids in length.

In certain embodiments of the invention, [G/D/E/-] is represented by an amino acid sequence selected from the group consisting of: G, D, E, and nothing.

In some embodiments of the invention, [N1] is represented by an amino acid sequence selected from the group consisting of: G, R, S, P, L, A, V, T, (G/P)(G/R/S/P/L/A/V/T), (R/S/L/A/V/T)(G/P), GG(G/R/S/P/L/A/V/T), G(R/S/P/L/A/V/T)G, (R/S/P/L/A/V/T)GG, and nothing.

In certain embodiments of the invention, [N2] is represented by an amino acid sequence selected from the group consisting of: G, R, S, P, L, A, V, T,

(G/P)(G/R/S/P/L/A/V/T), (R/S/L/A/V/T)(G/P), GG(G/R/S/P/L/A/V/T), G(R/S/P/L/A/V/T)G, (R/S/P/L/A/V/T)GG, and nothing.

In some embodiments of the invention, [DH] comprises a sequence selected from the group consisting of: IGHD3-10 reading frame 1, IGHD3-10 reading frame 2, IGHD3-10 reading frame 3, IGHD3-22 reading frame 2, IGHD6-19 reading frame 1, IGHD6-19 reading frame 2, IGHD6-13 reading frame 1, IGHD6-13 reading frame 2, IGHD3-03 reading frame 3, IGHD2-02 reading frame 2, IGHD2-02 reading frame 3, IGHD4-17 reading frame 2, IGHD1-26 reading frame 1, IGHD1-26 reading frame 3, IGHD5-5/5-18 reading frame 3, IGHD2-15 reading frame 2, and all possible N-terminal and C-terminal truncations of the above-identified IGHDs down to three amino acids.

In certain embodiments of the invention, [H3-JH] comprises a sequence selected from the group consisting of: AEYFQH, EYFQH, YFQH, FQH, QH, YWYFDL, WYFDL, YFDL, FDL, DL, AFDV, FDV, DV, YFDY, FDY, DY, NWFDS, WFDS, FDS, DS, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, MDV, and DV.

In some embodiments of the invention, the sequences represented by [G/D/E/-][N1][ext-DH][N2][H3-JH] comprise a sequence of about 3 to about 26 amino acids in length.

In certain embodiments of the invention, the sequences represented by [G/D/E/-][N1][ext-DH][N2][H3-JH] comprise a sequence of about 7 to about 23 amino acids in length.

In some embodiments of the invention, the library comprises about 10^7 to about 10^{10} sequences.

In certain embodiments of the invention, the library comprises about 10^7 sequences.

In some embodiments of the invention, the polynucleotide sequences of the libraries further comprise a 5’ polynucleotide sequence encoding a framework 3 (FRM3) region on the corresponding N-terminal end of the library sequence, wherein the FRM3 region comprises a sequence of about 1 to about 9 amino acid residues.

In certain embodiments of the invention , the FRM3 region comprises a sequence selected from the group consisting of CAR, CAK, and CAT.

In some embodiments of the invention, the polynucleotide sequences further comprise a 3’ polynucleotide sequence encoding a framework 4 (FRM4) region on the corresponding C-terminal end of the library sequence, wherein the FRM4 region comprises a sequence of about 1 to about 9 amino acid residues.

In certain embodiments of the invention, the library comprises a FRM4 region comprising a sequence selected from WGRG and WGQG.

In some embodiments of the invention, the polynucleotide sequences further comprise an FRM3 region coding for a corresponding polypeptide sequence comprising a sequence selected from the group consisting of CAR, CAK, and CAT; and an FRM4 region coding for a corresponding polypeptide sequence comprising a sequence selected from WGRG and WGQG.

In certain embodiments of the invention, the polynucleotide sequences further comprise 5' and 3' sequences which facilitate homologous recombination with a heavy chain chassis.

In some embodiments, the invention comprises a synthetic preimmune human antibody light chain library comprising polynucleotide sequences encoding human antibody kappa light chains represented by the formula:

[IGKV (1-95)][F/L/I/R/W/Y][JK].

In certain embodiments of the invention, [IGKV (1-95)] is selected from the group consisting of IGKV3-20 (1-95), IGKV1-39 (1-95), IGKV3-11 (1-95), IGKV3-15 (1-95), IGKV1-05 (1-95), IGKV4-01 (1-95), IGKV2-28 (1-95), IGKV 1-33 (1-95), IGKV1-09 (1-95), IGKV1-12 (1-95), IGKV2-30 (1-95), IGKV1-27 (1-95), IGKV1-16 (1-95), and truncations of said group up to and including position 95 according to Kabat.

In some embodiments of the invention, [F/L/I/R/W/Y] is an amino acid selected from the group consisting of F, L, I, R, W, and Y.

In certain embodiments of the invention, [JK] comprises a sequence selected from the group consisting of TFGQGTKVEIK and TFGGGT.

In some embodiments of the invention, the light chain library comprises a kappa light chain library.

In certain embodiments of the invention, the polynucleotide sequences further comprise 5' and 3' sequences which facilitate homologous recombination with a light chain chassis.

In some embodiments, the invention comprises a method for producing a synthetic preimmune human antibody CDRH3 library comprising 10^7 to 10^8 polynucleotide sequences, said method comprising:

a) selecting the CDRH3 polynucleotide sequences encoded by the CDRH3 sequences, as follows:

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by terminal deoxynucleotidyl transferase (TdT) and preferentially functionally expressed by human B cells}, followed by

{all possible N or C-terminal truncations of IGHD alone and all possible combinations of N and C-terminal truncations}, followed by

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by TdT and preferentially functionally expressed by human B cells}, followed by

{all possible N-terminal truncations of IGHJ, down to DXWG, wherein X is S, V, L, or Y}; and

b) synthesizing the CDRH3 library described in a) by chemical synthesis, wherein a synthetic preimmune human antibody CDRH3 library is produced.

In certain embodiments, the invention comprises a synthetic preimmune human antibody CDRH3 library comprising 10^7 to 10^{10} polynucleotide sequences representative of known human IGHD and IGHJ germline sequences encoding CDRH3, represented by the following formula:

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by terminal deoxynucleotidyl transferase (TdT) and preferentially functionally expressed by human B cells}, followed by

{all possible N or C-terminal truncations of IGHD alone and all possible combinations of N and C-terminal truncations}, followed by

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by TdT and preferentially functionally expressed by human B cells}, followed by

{all possible N-terminal truncations of IGHJ, down to DXWG, wherein X is S, V, L, or Y}.

In certain embodiments, the invention comprises a synthetic preimmune human antibody heavy chain variable domain library comprising 10^7 to 10^{10} polynucleotide sequences encoding human antibody heavy chain variable domains, said library comprising:

a) an antibody heavy chain chassis, and

b) a CDRH3 repertoire designed based on the human IGHD and IGHJ germline sequences, as follows:

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by terminal deoxynucleotidyl transferase (TdT) and preferentially functionally expressed by human B cells}, followed by

{all possible N or C-terminal truncations of IGHD alone and all possible combinations of N and C-terminal truncations}, followed by

{0 to 5 amino acids selected from the group consisting of fewer than ten of the amino acids preferentially encoded by TdT and preferentially functionally expressed by human B cells}, followed by

{all possible N-terminal truncations of IGHJ, down to DXWG, wherein X is S, V, L, or Y}.

In some embodiments of the invention, the synthetic preimmune human antibody heavy chain variable domain library is expressed as a full length chain selected from the group consisting of an IgG1 full length chain, an IgG2 full length chain, an IgG3 full length chain, and an IgG4 full length chain.

In certain embodiments of the invention, the human antibody heavy chain chassis is selected from the group consisting of IGHV4-34, IGHV3-23, IGHV5-51, IGHV1-69, IGHV3-30, IGHV4-39, IGHV1-2, IGHV1-18, IGHV2-5, IGHV2-70, IGHV3-7, IGHV6-1, IGHV1-46, IGHV3-33, IGHV4-31, IGHV4-4, IGHV4-61, and IGHV3-15.

In some embodiments of the invention, the synthetic preimmune human antibody heavy chain variable domain library comprises 10^7 to 10^{10} polynucleotide sequences encoding human antibody heavy chain variable domains, said library comprising:

a) an antibody heavy chain chassis, and

b) a synthetic preimmune human antibody CDRH3 library.

In some embodiments of the invention, the polynucleotide sequences are single-stranded coding polynucleotide sequences.

In certain embodiments of the invention, the polynucleotide sequences are single-stranded non-coding polynucleotide sequences.

In some embodiments of the invention, the polynucleotide sequences are double-stranded polynucleotide sequences.

In certain embodiments, the invention comprises a population of replicable cells with a doubling time of four hours or less, in which a synthetic preimmune human antibody repertoire is expressed.

In some embodiments of the invention, the population of replicable cells are yeast cells.

In certain embodiments, the invention comprises a method of generating a full-length antibody library comprising transforming a cell with a preimmune human antibody heavy chain variable domain library and a synthetic preimmune human antibody light chain library.

In some embodiments, the invention comprises a method of generating a full-length antibody library comprising transforming a cell with a preimmune human antibody heavy chain variable domain library and a synthetic preimmune human antibody light chain library.

In certain embodiments, the invention comprises a method of generating an antibody library comprising synthesizing polynucleotide sequences by split-pool DNA synthesis.

In some embodiments of the invention, the polynucleotide sequences are selected from the group consisting of single-stranded coding polynucleotide sequences, single-stranded non-coding polynucleotide sequences, and double-stranded polynucleotide sequences.

In certain embodiments, the invention comprises a synthetic full-length preimmune human antibody library comprising about 10^7 to about 10^{10} polynucleotide sequences representative of the sequence diversity and length diversity found in known heavy chain CDR3 sequences.

In certain embodiments, the invention comprises a method of selecting an antibody of interest from a human antibody library, comprising providing a synthetic preimmune human antibody CDRH3 library comprising a theoretical diversity of (N) polynucleotide sequences representative of the sequence diversity and length diversity found in known heavy chain CDR3 sequences, wherein the physical realization of that diversity is an actual library of a size at least 3(N), thereby providing a 95% probability that a single antibody of interest is present in the library, and selecting an antibody of interest.

In some embodiments of the invention, the theoretical diversity is about 10^7 to about 10^8 polynucleotide sequences.

EXAMPLES

This invention is further illustrated by the following examples which should not be construed as limiting. The contents of all references, patents and published patent applications cited throughout this application are hereby incorporated by reference.

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.*, yeast expression, cell-free expression, phage display, ribosome display, and PROFUSION^{TM}), 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.*, Nature Biotechnology, 1997, 15: 553; Border *et al.*, Methods Enzymol., 2000, 328: 430; ribosome display as described by Pluckthun *et al.* in U.S. Patent No. 6,348,315, and Profusion^{TM} 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. Each of the references cited in this paragraph is incorporated by reference in its entirety.

Further details regarding antibody sequence analysis using Kabat conventions and programs to screen aligned nucleotide and amino acid sequences may be found, *e.g.*, in Johnson *et al.*, Methods Mol. Biol., 2004, 248: 11; Johnson *et al.*, Int. Immunol., 1998, 10: 1801; Johnson *et al.*, Methods Mol. Biol., 1995, 51: 1; Wu *et al.*, Proteins, 1993, 16: 1; and Martin, Proteins, 1996, 25: 130. Each of the references cited in this paragraph is incorporated by reference in its entirety.

Further details regarding antibody sequence analysis using Chothia conventions may be found, *e.g.*, in Chothia *et al.*, J. Mol. Biol., 1998, 278: 457; Morea *et al.*, Biophys. Chem., 1997, 68: 9; Morea *et al.*, J. Mol. Biol., 1998, 275: 269; Al-Lazikani *et al.*, J. Mol. Biol., 1997, 273: 927. Barre *et al.*, Nat. Struct. Biol., 1994, 1: 915; Chothia *et al.*, J. Mol. Biol., 1992, 227: 799; Chothia *et al.*, Nature, 1989, 342: 877; and Chothia *et al.*, J. Mol. Biol., 1987, 196: 901. Further analysis of CDRH3 conformation may be found in Shirai *et al.*, FEBS Lett., 1999, 455: 188 and Shirai *et al.*, FEBS Lett., 1996, 399: 1. Further details regarding Chothia analysis are described, for example, in Chothia *et al.*, Cold Spring Harb. Symp. Quant Biol., 1987, 52: 399. Each of the references cited in this paragraph is incorporated by reference in its entirety.

Further details regarding CDR contact considerations are described, for example, in MacCallum *et al.*, J. Mol. Biol., 1996, 262: 732, incorporated by reference in its entirety.

Further details regarding the antibody sequences and databases referred to herein are found, *e.g.*, in Tomlinson *et al.*, J. Mol. Biol., 1992, 227: 776, VBASE2 (Retter *et al.*, Nucleic Acids Res., 2005, 33: D671); BLAST (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). Each of the references cited in this paragraph is incorporated by reference in its entirety.

Example 1: Design of an Exemplary VH Chassis Library

This example demonstrates the selection and design of exemplary, non-limiting VH chassis sequences of the invention. VH chassis sequences were selected by examining collections of human IGHV germline sequences (Scaviner *et al.*, Exp. Clin. Immunogenet., 1999, 16: 234; Tomlinson *et al.*, J. Mol. Biol., 1992, 227: 799; Matsuda *et al.*, J. Exp. Med., 1998, 188: 2151, each incorporated by reference in its entirety). As discussed in the Detailed Description, as well as below, a variety of criteria can be used to select VH chassis sequences, from these data sources or others, for inclusion in the library.

Table 3 (adapted from information provided in Scaviner *et al.*, Exp. Clin. Immunogenet., 1999, 16: 234; Matsuda *et al.,* J. Exp. Med., 1998, 188: 2151; and Wang *et al.* Immunol. Cell. Biol., 2008, 86: 111, each incorporated by reference in its entirety) lists the CDRH1 and CDRH2 length, the canonical structure and the estimated relative occurrence in peripheral blood, for the proteins encoded by each of the human IGHV germline sequences.

Table 3. IGHV Characteristics and Occurrence in Antibodies from Peripheral Blood

IGHV Germline	Length of CDRH1	Length of CDRH2	Canonical Structures[1]	Estimated Relative Occurrence in Peripheral Blood[2]
IGHV1-2	5	17	1-3	37
IGHV1-3	5	17	1-3	15
IGHV1-8	5	17	1-3	13
IGHV1-18	5	17	1-2	25
IGHV1-24	5	17	1-U	5
IGHV1-45	5	17	1-3	0
IGHV1-46	5	17	1-3	25
IGHV1-58	5	17	1-3	2
IGHV1-69	5	17	1-2	58
IGHV2-5	7	16	3-1	10
IGHV2-26	7	16	3-1	9
IGHV2-70	7	16	3-1	13
IGHV3-7	5	17	1-3	26
IGHV3-9	5	17	1-3	15
IGHV3-11	5	17	1-3	13
IGHV3-13	5	16	1-1	3
IGHV3-15	5	19	1-4	14
IGHV3-20	5	17	1-3	3
IGHV3-21	5	17	1-3	19
IGHV3-23	5	17	1-3	80
IGHV3-30	5	17	1-3	67
IGHV3-33	5	17	1-3	28
IGHV3-43	5	17	1-3	2
IGHV3-48	5	17	1-3	21
IGHV3-49	5	19	1-U	8
IGHV3-53	5	16	1-1	7
IGHV3-64	5	17	1-3	2
IGHV3-66	5	17	1-3	3
IGHV3-72	5	19	1-4	2
IGHV3-73	5	19	1-4	3
IGHV3-74	5	17	1-3	14
IGHV4-4	5	16	1-1	33
IGHV4-28	6	16	2-1	1

IGHV4-31	7	16	3-1	25
IGHV4-34	5	16	1-1	125
IGHV4-39	7	16	3-1	63
IGHV4-59	5	16	1-1	51
IGHV4-61	7	16	3-1	23
IGHV4-B	6	16	2-1	7
IGHV5-51	5	17	1-2	52
IGHV6-1	7	18	3-5	26
IGHV7-4-1	5	17	1-2	8

[1] Adapted from Chothia *et al.*, J. Mol. Biol., 1992, 227: 799

[2] Adapted from Table S1 of Wang *et al.*, Immunol. Cell. Biol., 2008, 86: 111

In the currently exemplified library, 17 germline sequences were chosen for representation in the VH chassis of the library (Table 4). As described in more detail below, these sequences were selected based on their relatively high representation in the peripheral blood of adults, with consideration given to the structural diversity of the chassis and the representation of particular germline sequences in antibodies used in the clinic. These 17 sequences account for about 76% of the total sample of heavy chain sequences used to derive the results of Table 4. As outlined in the Detailed Description, these criteria are non-limiting, and one of ordinary skill in the art will readily recognize that a variety of other criteria can be used to select the VH chassis sequences, and that the invention is not limited to a library comprising the 17 VH chassis genes presented in Table 4.

Table 4. VH Chassis Selected for Use in the Exemplary Library

VH Chassis	Relative Occurrence	Length of CDRH1	Length of CDRH2	Comment
VH1-2	37	5	17	Among highest usage for VH1 family
VH1-18	25	5	17	Among highest usage for VH1 family
VH1-46	25	5	17	Among highest usage for VH1 family
VH1-69	58	5	17	Highest usage for VH1 family. The four chosen VH1 chassis represent about 80% of the VH1 repertoire.
VH3-7	26	5	17	Among highest usage in VH3 family
VH3-15	14	5	19	Not among highest usage, but it has unique structure (H2 of length 19). Highest occurrence among those with such structure.
VH3-23	80	5	17	Highest usage in VH3

				family.
VH3-30	67	5	17	Among highest usage in VH3 family
VH3-33	28	5	17	Among highest usage in VH3 family
VH3-48	21	5	17	Among highest usage in VH3 family. The six chosen VH3 chassis account for about 70% of the VH3 repertoire.
VH4-31	25	7	16	Among highest usage in VH4 family
VH4-34	125	5	16	Highest usage in VH4 family
VH4-39	63	7	16	Among highest usage in VH4 family
VH4-59	51	5	16	Among highest usage in VH4 family
VH4-61	23	7	16	Among highest usage in VH4 family
VH4-B	7	6	16	Not among highest usage in VH4 family, but has unique structure (H1 of length 6). The 6 chosen VH4 chassis account for close to 90% of the VH4 family repertoire
VH5-51	52	5	17	High usage

In this particular embodiment of the library, VH chassis derived from sequences in the IGHV2, IGHV6 and IGHV7 germline families were not included. As described in the Detailed Description, this exemplification is not meant to be limiting, as, in some embodiments, it may be desirable to include one or more of these families, particularly as clinical information on antibodies with similar sequences becomes available, to produce libraries with additional diversity that is potentially unexplored, or to study the properties and potential of these IGHV families in greater detail. The modular design of the library of the present invention readily permits the introduction of these, and other, VH chassis sequences. The amino acid sequences of the VH chassis utilized in this particular embodiment of the library, which are derived from the IGHV germline sequences, are presented in Table 5. The details of the derivation procedures are presented below.

Table 5. Amino Acid Sequences for VH Chassis Selected for Inclusion in the Exemplary Library

Chassis	**SEQ ID NO:**	**FRM1**	**CDRH1**	**FRM2**	**CDRH2**	**FRM3**
VH1-2		`QVQLVQSGAEVKKPGA`	`GYYMH`	`WVRQAPGQGLEWMG`	`WINPNSGGTNYAQK`	`RVTMTRDTSISTAYMELSRL`

		SVKVSCKASGYTFT			FQG	RSDDTAVYYCAR
VH1-18		QVQLVQSGAEVKKPGASVKVSCKASGYTFT	SYGIS	WVRQAPGQGLEWMG	WISAYNGNTNYAQKLQG	RVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR
VH1-46		QVQLVQSGAEVKKPGASVKVSCKASGYTFT	SYYMH	WVRQAPGQGLEWMG	IINPSGGSTSYAQKFQG	RVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR
VH1-69		QVQLVQSGAEVKKPGSSVKVSCKASGGTFS	SYAIS	WVRQAPGQGLEWMG	GIIPIFGTANYAQKFQG	RVTITADKSTSTAYMELSSLRSEDTAVYYCAR
VH3-7		EVQLVESGGGLVQPGGSLRLSCAASGFTFS	SYWMS	WVRQAPGKGLEWVA	NIKQDGSEKYYVDSVKG	RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
VH3-15[1]		EVQLVESGGGLVKPGGSLRLSCAASGFTFS	NAWMS	WVRQAPGKGLEWVG	RIKSKTDGGTTDYAAPVKG	RFTISRDDSKNTLYLQMNSL**RA**EDTAVYYC**AR**
VH3-23		EVQLLESGGGLVQPGGSLRLSCAASGFTFS	SYAMS	WVRQAPGKGLEWVS	AISGSGGSTYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK
VH3-30		QVQLVESGGGVVQPGRSLRLSCAASGFTFS	SYGMH	WVRQAPGKGLEWVA	VISYDGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
VH3-33		QVQLVESGGGVVQPGRSLRLSCAASGFTFS	SYGMH	WVRQAPGKGLEWVA	VIWYDGSNKYYADSVKG	RFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR
VH3-48		EVQLVESGGGLVQPGGSLRLSCAASGFTFS	SYSMN	WVRQAPGKGLEWVS	YISSSSSTIYYADSVKG	RFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR
VH4-31		QVQLQESGPGLVKPSQTLSLTCTVSGGSIS	SGGYYWS	WIRQHPGKGLEWIG	YIYYSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR

VH4-34[2]		QVQLQQWGAGLLKPSETLSLTCAVYGGSFS	GYYWS	WIRQPPGKGLEWIG	EI**DHS**GSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
VH4-39		QLQLQESGPGLVKPSETLSLTCTVSGGSIS	SSSYYWG	WIRQPPGKGLEWIG	SIYYSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
VH4-59		QVQLQESGPGLVKPSETLSLTCTVSGGSIS	SYYWS	WIRQPPGKGLEWIG	YIYYSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
VH4-61		QVQLQESGPGLVKPSETLSLTCTVSGGSVS	SGSYYWS	WIRQPPGKGLEWIG	YIYYSGSTNYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
VH4-B		QVQLQESGPGLVKPSETLSLTCAVSGYSIS	SGYYWG	WIRQPPGKGLEWIG	SIYHSGSTYYNPSLKS	RVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
VH5-51		EVQLVQSGAEVKKPGESLKISCKGSGYSFT	SYWIG	WVRQMPGKGLEWMG	IIYPGDSDTRYSPSFQG	QVTISADKSISTAYLQWSSLKASDTAVYYCAR

[1] The original KT sequence in VH3-15 was mutated to RA (bold/underlined) and TT to AR (bold/underlined), in order to match other VH3 family members selected for inclusion in the library. The modification to RA was made so that no unique sequence stretches of up to about 20 amino acids are created. Without being bound by theory, this modification is expected to reduce the odds of introducing novel T-cell epitopes in the VH3-15-derived chassis sequence. The avoidance of T cell epitopes is an additional criterion that can be considered in the design of certain libraries of the invention.

[2] The original NHS motif in VH4-34 was mutated to DHS, in order to remove a possible N-linked glycosylation site in CDR-H2. In certain embodiments of the invention, for example, if the library is transformed into yeast, this may prevent unwanted N-linked glycosylation.

Table 5 provides the amino acid sequences of the seventeen chassis. In nucleotide space, most of the corresponding germline nucleotide sequences include two additional nucleotides on the 3’ end (*i.e.,* two-thirds of a codon). In most cases, those two nucleotides are GA. In many cases, nucleotides are added to the 3’ end of the IGHV-derived gene segment *in vivo*, prior to recombination with the IGHD gene segment. Any additional nucleotide would make the resulting codon encode one of the following two amino acids: Asp (if the codon is GA**C** or GA**T**) or Glu (if the codon is

GAA or GAG). One, or both, of the two 3'-terminal nucleotides may also be deleted in the final rearranged heavy chain sequence. If only the A is deleted, the resulting amino acid is very frequently a G. If both nucleotides are deleted, this position is "empty," but followed by a general V-D addition or an amino acid encoded by the IGHD gene. Further details are presented in Example 5. This first position, after the CAR or CAK motif at the C-terminus of FRM3 (Table 5), is designated the "tail." In the currently exemplified embodiment of the library, this residue may be G, D, E, or nothing. Thus, adding the tail to any chassis enumerated above (Table 5) can produce one of the following four schematic sequences, wherein the residue following the VH chassis is the tail:

(1) [VH_Chassis]-[G]
(2) [VH_Chassis]-[D]
(3) [VH_Chassis]-[E]
(4) [VH_Chassis]

These structures can also be represented in the format:

[VH_Chassis]-[G/D/E/-],

wherein the hyphen symbol (-) indicates an empty or null position.

Using the CDRH3 numbering system defined in the Definitions section, the above sequences could be denoted to have amino acid 95 as G, D, or E, for instances (1), (2), and (3), respectively, while the sequence of instance 4 would have no position 95, and CDRH3 proper would begin at position 96 or 97.

In some embodiments of the invention, VH3-66, with canonical structure 1-1 (five residues in CDRH1 and 16 for CDRH2) may be included in the library. The inclusion of VH3-66 may compensate for the removal of other chassis from the library, which may not express well in yeast under some conditions (*e.g.,* VH4-34 and VH4-59).

Example 2: Design of VH Chassis Variants with Variation Within CDRH1 and CDRH2

This example demonstrates the introduction of further diversity into the VH chassis by creating mutations in the CDRH1 and CDRH2 regions of each chassis shown in Example 1. The following approach was used to select the positions and nature of the amino acid variation for each chassis: First, the sequence identity between rearranged human heavy chain antibody sequences was analyzed (Lee *et al.,* Immunogenetics, 2006, 57: 917; Jackson *et al.*, J. Immunol. Methods, 2007, 324: 26) and they were classified by the origin of their respective IGHV germline sequence. As an illustrative example, about 200 sequences in the data set exhibited greatest identity to the IGHV1-69 germline, indicating that they were likely to have been derived from IGHV1-69. Next, the occurrence of amino acid residues at each position within the CDRH1 and CDRH2 segments, in each germline family selected in Example 1 was determined. For VH1-69, these occurrences are illustrated in Tables 6 and 7. Second, neutral and/or smaller amino acid residues were favored, where possible, as replacements. Without being bound by theory, the rationale for the choice of these amino acid residues is the desire to provide a more flexible and less sterically hindered context for the display of a diversity of CDR sequences.

Table 6. Occurrence of Amino Acid Residues at Each Position Within IGHV1-69-derived CDRH1 Sequences

	31	**32**	**33**	**34**	**35**
	S	**Y**	**A**	**I**	**S**
A	1	0	**129**	0	0
C	0	1	0	0	2
D	0	5	1	0	0
E	0	0	0	0	0
F	0	9	1	8	0
G	0	0	**24**	0	3
H	2	11	0	0	4
I	2	0	0	**159**	1
K	3	0	0	0	0
L	0	**10**	2	5	0
M	1	0	0	0	0
N	**21**	2	2	0	**27**
P	0	0	1	0	0
Q	1	1	0	0	5
R	9	0	0	0	1
S	**133**	3	7	0	**129**
T	12	1	**10**	0	12
V	0	0	7	13	0
W	0	0	0	0	0
Y	0	**142**	1	0	1

Table 7. Occurrence of Amino Acid Residues at Each Position Within IGHV1-69-derived CDRH2 Sequences

	50	**51**	**52**	**52A**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**	**64**	**65**
	G	**I**	**I**	**P**	**I**	**F**	**G**	**T**	**A**	**N**	**Y**	**A**	**Q**	**K**	**F**	**Q**	**G**
A	0	0	**7**	0	2	0	4	3	**132**	0	0	**178**	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	0	0	0	0	0	11	0	1	21	0	0	0	2	0	0	12
E	2	0	0	0	0	0	4	0	0	2	0	1	1	4	0	2	0
F	0	1	0	1	7	**119**	0	0	0	0	0	0	0	0	**180**	0	0
G	**135**	0	1	0	0	0	**155**	0	3	1	0	0	0	0	0	0	**173**
H	0	0	0	0	1	0	0	0	0	4	4	0	3	0	0	4	0
I	0	**166**	**159**	0	**132**	2	0	34	0	2	1	0	0	0	0	0	0
K	1	0	0	0	0	0	0	4	1	5	0	0	2	**156**	0	3	0
L	0	1	2	0	**16**	37	0	1	0	0	0	0	0	0	3	2	0
M	0	6	2	0	9	1	0	3	1	0	0	0	0	0	0	0	0
N	0	0	1	0	2	0	5	0	0	**132**	1	0	0	8	0	0	0
P	0	2	0	**181**	1	3	0	0	15	0	0	3	6	0	0	0	0
Q	0	0	0	0	0	0	1	0	1	0	0	0	**173**	2	0	**164**	0
R	**44**	0	0	0	0	0	1	4	0	3	0	0	0	13	0	9	0
S	1	0	1	1	2	6	3	5	8	**7**	0	2	0	0	1	0	0
T	1	1	7	2	2	1	0	**127**	15	8	3	1	0	0	0	0	0
V	0	8	5	0	11	4	0	4	8	0	0	0	0	0	0	0	0
W	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	11	1	0	0	0	**176**	0	0	0	1	1	0

The original germline sequence is provided in the second row of the tables, in bold font, beneath the residue number (Kabat system). The entries in the table indicate the number of times a given amino acid residue (first column) is observed at the indicated CDRH1 (Table 6) or CDRH2 (Table 7) position. For example, at position 33 the amino acid type G (glycine) is observed 24 times in the set of IGHV1-69-based sequences that were examined. Thus, applying the criteria above, variants were constructed with N at position 31, L at position 32 (H can be charged, under some conditions), G and T at position 33, no variants at position 34 and N at position 35, resulting in the following VH1-69 chassis CDRH1 single-amino acid variant sequences:

NYAIS (SEQ ID NO: ___)
S**L**AIS (SEQ ID NO: ___)
SY**G**IS (SEQ ID NO: ___)
SY**T**IS (SEQ ID NO: ___)
SYAI**N** (SEQ ID NO: ___)

Similarly, the analysis that produced Table 7 provided a basis for choosing the following single-amino acid variant sequences for VH1-69 chassis CDRH2s:

SIIPIFGTANYAQKFQG (SEQ ID NO: ___)
GI**A**PIFGTANYAQKFQG (SEQ ID NO: ___)
GIIPI**L**GTANYAQKFQG (SEQ ID NO: ___)
GIIPIFGTA**S**YAQKFQG (SEQ ID NO: ___)

A similar approach was used to design and construct variants of the other selected chassis; the resulting CDRH1 and CDRH2 variants for each of the exemplary chassis are provided in Table 8. One of ordinary skill in the art will readily recognize that the methods described herein can be applied to create variants of other VH chassis and VL chassis.

Table 8. VH Chassis Variants

Chassis	CDRH1	SEQ ID NO:	CDRH2	SEQ ID NO:	Chassis	CDRH1	SEQ ID NO:	CDRH2	SEQ ID NO:
1-18.0	SYGIS		WISAYNGNTNYAQKLQG		**3-48.0**	SYSMN		YISSSSSTIYYADSVKG	
1-18.1	**N**YGIS		WISAYNGNTNYAQKLQG		**3-48.1**[1]	**N**YSMN		YISSSSSTIYYADSVKG	
1-18.2	S**N**GIS		WISAYNGNTNYAQKLQG		**3-48.2**	**I**YSMN		YISSSSSTIYYADSVKG	
1-18.3	SY**A**IS		WISAYNGNTNYAQKLQG		**3-48.3**	S**N**SMN		YISSSSSTIYYADSVKG	
1-18.4	SYGI**T**		WISAYNGNTNYAQKLQG		**3-48.4**	SY**E**MN		YISSSSSTIYYADSVKG	
1-18.5	SYGI**H**		WISAYNGNTNYAQKLQG		**3-48.5**	SY**N**MN		YISSSSSTIYYADSVKG	
1-18.6	SYGIS		**S**ISAYNGNTNYAQKLQG		**3-48.6**	SYSM**T**		YISSSSSTIYYADSVKG	
1-18.7	SYGIS		WIS**T**YNGNTNYAQKLQG		**3-48.7**	SYSMN		TISSSSSTIYYADSVKG	
1-18.8	SYGIS		WIS**P**YNGNTNYAQKLQG		**3-48.8**	SYSMN		YISGSSSTIYYADSVKG	
1-18.9	SYGIS		WIS**A**YNGNTYYAQKLQG		**3-48.9**	SYSMN		YISSSSSTI**L**YADSVKG	

1-2.0	GYYMH		WINPNSGGT NYAQKFQG		**3-7.0**	SYWMS		NIKQDGSEK YYVDSVKG	
1-2.1	**D**YYMH		WINPNSGGT NYAQKFQG		**3-7.1**	**T**YWMS		NIKQDGSEK YYVDSVKG	
1-2.2	**R**YYMH		WINPNSGGT NYAQKFQG		**3-7.2**	**N**YWMS		NIKQDGSEK YYVDSVKG	
1-2.3	G**S**YMH		WINPNSGGT NYAQKFQG		**3-7.3**	S**S**WMS		NIKQDGSEK YYVDSVKG	
1-2.4	GY**S**MH		WINPNSGGT NYAQKFQG		**3-7.4**	SY**G**MS		NIKQDGSEK YYVDSVKG	
1-2.5	GYYM**Q**		WINPNSGGT NYAQKFQG		**3-7.5**	SYWM**T**		NIKQDGSEK YYVDSVKG	
1-2.6	GYYMH		**S**INPNSGGT NYAQKFQG		**3-7.6**	SYWMS		**S**IKQDGSEK YYVDSVKG	
1-2.7	GYYMH		WINP**S**SGGT NYAQKFQG		**3-7.7**	SYWMS		NI**N**QDGSEK YYVDSVKG	
1-2.8	GYYMH		WINPNSGGT **K**YAQKFQG		**3-7.8**	SYWMS		NIK**S**DGSEK YYVDSVKG	
1-2.9	GYYMH		WINPNSGGT **S**YAQKFQG		**3-7.9**	SYWMS		NIKQDGSEK **Q**YVDSVKG	
1-46.0	SYYMH		IINPSGGST SYAQKFQG		**4-31.0**	SGGYYWS		YIYYSGSTY YNPSLKS	

1-46.1	**N**YYMH		IINPSGGSTSYAQKFQG		**4-31.1**	SG**S**YYWS		YIYYSGSTYYNPSLKS	
1-46.2	S**S**YMH		IINPSGGSTSYAQKFQG		**4-31.2**	SG**T**YYWS		YIYYSGSTYYNPSLKS	
1-46.3	SY**S**MH		IINPSGGSTSYAQKFQG		**4-31.3**	SGG**T**YWS		YIYYSGSTYYNPSLKS	
1-46.4	SYY**I**H		IINPSGGSTSYAQKFQG		**4-31.4**	SGGY**S**WS		YIYYSGSTYYNPSLKS	
1-46.5	SYYM**V**		IINPSGGSTSYAQKFQG		**4-31.5**	SGGYYWS		**S**IYYSGSTYYNPSLKS	
1-46.6	SYYM**S**		IINPSGGSTSYAQKFQG		**4-31.6**	SGGYYWS		**N**IYYSGSTYYNPSLKS	
1-46.7	SYYMH		**V**INPSGGSTSYAQKFQG		**4-31.7**	SGGYYWS		YIYYSG**N**TYYNPSLKS	
1-46.8	SYYMH		IINP**G**GGSTSYAQKFQG		**4-31.8**	SGGYYWS		YIYYSGST**S**YNPSLKS	
1-46.9	SYYMH		IINPSGGST**T**YAQKFQG		**4-31.9**	SGGYYWS		YIYYSGST**V**YNPSLKS	
1-69.0	SYAIS		GIIPIFGTANYAQKFQG		**4-34.0**	GYYWS		EIDHSGSTNYNPSLKS	
1-69.1	**N**YAIS		GIIPIFGTANYAQKFQG		**4-34.1**	**D**YYWS		EIDHSGSTNYNPSLKS	

1-69.2	S**L**AIS		GIIPIFGTA NYAQKFQG		**4-34.2**	GYYW**T**		EIDHSGSTN YNPSLKS	
1-69.3	SY**G**IS		GIIPIFGTA NYAQKFQG		**4-34.3**	GYYWS		**D**IDHSGSTN YNPSLKS	
1-69.4	SY**T**IS		GIIPIFGTA NYAQKFQG		**4-34.4**	GYYWS		EI**S**HSGSTN YNPSLKS	
1-69.5	SYAI**N**		GIIPIFGTA NYAQKFQG		**4-34.5**	GYYWS		EID**Q**SGSTN YNPSLKS	
1-69.6	SYAIS		**S**IIPIFGTA NYAQKFQG		**4-34.6**	GYYWS		EIDH**G**GSTN YNPSLKS	
1-69.7	SYAIS		GI**A**PIFGTA NYAQKFQG		**4-34.7**	GYYWS		EIDHSG**N**TN YNPSLKS	
1-69.8	SYAIS		GIIPI**L**GTA NYAQKFQG		**4-34.8**	GYYWS		EIDHSGST**S** YNPSLKS	
1-69.9	SYAIS		GIIPIFGTA **S**YAQKFQG		**4-34.9**	GYYWS		EIDHSGST**D** YNPSLKS	
3-15.0	NAWMS		RIKSKTDGG TTDYAAPVK G		**4-39.0**	SSSYYWG		SIYYSGSTY YNPSLKS	
3-15.1	**K**AWMS		RIKSKTDGG TTDYAAPVK G		**4-39.1**	**T**SSYYWG		SIYYSGSTY YNPSLKS	
3-15.2	**D**AWMS		RIKSKTDGG		**4-39.2**	S**N**SYYWG		SIYYSGSTY	

			TTDYAAPVKG					YNPSLKS	
3-15.3	NA**L**MS		RIKSKTDGGTTDYAAPVKG		**4-39.3**	SS**D**YYWG		SIYYSGSTYYNPSLKS	
3-15.4	NA**A**MS		RIKSKTDGGTTDYAAPVKG		**4-39.4**	SS**N**YYWG		SIYYSGSTYYNPSLKS	
3-15.5	NAWM**N**		RIKSKTDGGTTDYAAPVKG		**4-39.5**	SS**R**YYWG		SIYYSGSTYYNPSLKS	
3-15.6	NAWMS		**S**IKSKTDGGTTDYAAPVKG		**4-39.6**	SSSY**A**WG		SIYYSGSTYYNPSLKS	
3-15.7	NAWMS		RIKS**T**TDGGTTDYAAPVKG		**4-39.7**	SSSYYWG		**N**IYYSGSTYYNPSLKS	
3-15.8	NAWMS		RIKSK**A**DGGTTDYAAPVKG		**4-39.8**	SSSYYWG		SI**S**YSGSTYYNPSLKS	
3-15.9	NAWMS		RIKSKTDGGTT**G**YAAPVKG		**4-39.9**	SSSYYWG		SIYYSGST**S**YNPSLKS	
3-23.0	SYAMS		AISGSGGSTYYADSVKG		**4-59.0**	SYYWS		YIYYSGST**N**YNPSLKS	

3-23.1	**N**YAMS		AISGSGGSTYYADSVKG		**4-59.1**	**T**YYWS		YIYYSGSTNYNPSLKS	
3-23.2	**T**YAMS		AISGSGGSTYYADSVKG		**4-59.2**	S**S**YWS		YIYYSGSTNYNPSLKS	
3-23.3	S**S**AMS		AISGSGGSTYYADSVKG		**4-59.3**	SY**S**WS		YIYYSGSTNYNPSLKS	
3-23.4	SYAMS		**G**ISGSGGSTYYADSVKG		**4-59.4**	SYYWS		**F**IYYSGSTNYNPSLKS	
3-23.5	SYAMS		**S**ISGSGGSTYYADSVKG		**4-59.5**	SYYWS		**H**IYYSGSTNYNPSLKS	
3-23.6	SYAMS		**T**ISGSGGSTYYADSVKG		**4-59.6**	SYYWS		**S**IYYSGSTNYNPSLKS	
3-23.7	SYAMS		**V**ISGSGGSTYYADSVKG		**4-59.7**	SYYWS		YIY**S**SGSTNYNPSLKS	
3-23.8	SYAMS		AIS**A**SGGSTYYADSVKG		**4-59.8**	SYYWS		YIYYSGST**D**YNPSLKS	
3-23.9	SYAMS		AISGSGGST**S**YADSVKG		**4-59.9**	SYYWS		YIYYSGST**T**YNPSLKS	
3-30.0	SYGMH		VISYDGSNKYYADSVKG		**4-61.0**	SGSYYWS		YIYYSGSTNYNPSLKS	
3-30.1	NYGMH		VISYDGSNKYYADSVKG		**4-61.1**	SG**G**YYWS		YIYYSGSTNYNPSLKS	

3-30.2	SY**A**MH		VISYDGSNK YYADSVKG		**4-61.2**	SG**N**YYWS		YIYYSGSTN YNPSLKS	
3-30.3	SYG**F**H		VISYDGSNK YYADSVKG		**4-61.3**	SGS**S**YWS		YIYYSGSTN YNPSLKS	
3-30.4	SYGMH		**F**ISYDGSNK YYADSVKG		**4-61.4**	SGSY**S**WS		YIYYSGSTN YNPSLKS	
3-30.5	SYGMH		**L**ISYDGSNK YYADSVKG		**4-61.5**	SGSYYW**T**		YIYYSGSTN YNPSLKS	
3-30.6	SYGMH		VIS**S**DGSNK YYADSVKG		**4-61.6**	SGSYYWS		**R**IYYSGSTN YNPSLKS	
3-30.7	SYGMH		VISYDG**N**NK YYADSVKG		**4-61.7**	SGSYYWS		**S**IYYSGSTN YNPSLKS	
3-30.8	SYGMH		VISYDGS**I**K YYADSVKG		**4-61.8**	SGSYYWS		YIY**T**SGSTN YNPSLKS	
3-30.9	SYGMH		VISYDGSN**Q** YYADSVKG		**4-61.9**	SGSYYWS		YIYYSGST**S** YNPSLKS	
3-33.0	SYGMH		VIWYDGSNK YYADSVKG		**4-B.0**	SGYYWG		SIYHSGSTY YNPSLKS	
3-33.1	**T**YGMH		VIWYDGSNK YYADSVKG		**4-B.1**	S**A**YYWG		SIYHSGSTY YNPSLKS	
3-33.2	**N**YGMH		VIWYDGSNK YYADSVKG		**4-B.2**	SG**S**YWG		SIYHSGSTY YNPSLKS	

3-33.3	S**S**GMH		VIWYDGSNKYYADSVKG		**4-B.3**	SGY**N**WG		SIYHSGSTYYNPSLKS	
3-33.4	SY**A**MH		VIWYDGSNKYYADSVKG		**4-B.4**	SGYYW**A**		SIYHSGSTYYNPSLKS	
3-33.5	SYGM**N**		VIWYDGSNKYYADSVKG		**4-B.5**	SGYYWG		**T**IYHSGSTYYNPSLKS	
3-33.6	SYGMH		**L**IWYDGSNKYYADSVKG		**4-B.6**	SGYYWG		S**S**YHSGSTYYNPSLKS	
3-33.7	SYGMH		**F**IWYDGSNKYYADSVKG		**4-B.7**	SGYYWG		SIYHSG**N**TYYNPSLKS	
3-33.8	SYGMH		VIWYDGSNK**S**YADSVKG		**4-B.8**	SGYYWG		SIYHSGST**N**YNPSLKS	
3-33.9	SYGMH		VIWYDGSNK**G**YADSVKG		**4-B.9**	SGYYWG		SIYHSGST**G**YNPSLKS	
					5-51.0	SYWIG		IIYPGDSDTRYSPSFQG	
					5-51.1	**T**YWIG		IIYPGDSDTRYSPSFQG	
					5-51.2	**N**YWIG		IIYPGDSDTRYSPSFQG	
					5-51.3	S**N**WIG		IIYPGDSDTRYSPSFQG	

					5-51.4	SY**Y**IG		IIYPGDSDT RYSPSFQG	
					5-51.5	SYWI**S**		IIYPGDSDT RYSPSFQG	
					5-51.6	SYWIG		**S**IYPGDSDT RYSPSFQG	
					5-51.7	SYWIG		IIYP**A**DSDT RYSPSFQG	
					5-51.8	SYWIG		IIYPGDS**S**T RYSPSFQG	
					5-51.9	SYWIG		IIYPGDSDT **T**YSPSFQG	

---

[1] Contains an N-linked glycosylation site which can be removed, if desired, as described herein.

As specified in the Detailed Description, other criteria can be used to select which amino acids are to be altered and the identity of the resulting altered sequence. This is true for any heavy chain chassis sequence, or any other sequence of the invention. The approach outlined above is meant for illustrative purposes and is non-limiting.

Example 3: Design of an Exemplary VK Chassis Library

This example describes the design of an exemplary VK chassis library. One of ordinary skill in the art will recognize that similar principles may be used to design a Vλ library, or a library containing both VK and Vλ chassis. Design of a Vλ chassis library is presented in Example 4.

As was previously demonstrated in Example 1, for IGHV germline sequences, the sequence characteristics and occurrence of human IGKV germline sequences in antibodies from peripheral blood were analyzed. The data are presented in Table 9.

Table 9. IGKV Gene Characteristics and Occurrence in Antibodies from Peripheral Blood

IGKV Gene	Alternative Names	CDRL1 Length	CDRL2 Length	Canonical Structures[1]	Estimated Relative Occurrence in Peripheral Blood[2]
IGKV1-05	L12	11	7	2-1-(U)	**69**
IGKV1-06	L11	11	7	2-1-(1)	14
IGKV1-08	L9	11	7	2-1-(1)	9
IGKV1-09	L8	11	7	2-1-(1)	**24**
IGKV1-12	L5, L19	11	7	2-1-(1)	**32**
IGKV1-13	L4, L18	11	7	2-1-(1)	13
IGKV1-16	L1	11	7	2-1-(1)	15
IGKV1-17	A30	11	7	2-1-(1)	**34**
IGKV1-27	A20	11	7	2-1-(1)	**27**
IGKV1-33	O8, O18	11	7	2-1-(1)	**43**
IGKV1-37	O14, O4	11	7	2-1-(1)	3
IGKV1-39	O2, O12	11	7	2-1-(1)	**147**
IGKV1D-16	L15	11	7	2-1-(1)	6
IGKV1D-17	L14	11	7	2-1-(1)	1
IGKV1D-43	L23	11	7	2-1-(1)	1
IGKV1D-8	L24	11	7	2-1-(1)	1
IGKV2-24	A23	16	7	4-1-(1)	8
IGKV2-28	A19, A3	16	7	4-1-(1)	**62**
IGKV2-29	A18	16	7	4-1-(1)	6
IGKV2-30	A17	16	7	4-1-(1)	**30**
IGKV2-40	O1, O11	17	7	3-1-(1)	3
IGKV2D-26	A8	16	7	4-1-(1)	0

IGKV2D-29	A2	16	7	4-1-(1)	**20**
IGKV2D-30	A1	16	7	4-1-(1)	4
IGKV3-11	L6	11	7	2-1-(1)	**87**
IGKV3-15	L2	11	7	2-1-(1)	**53**
IGKV3-20	A27	12	7	6-1-(1)	**195**
IGKV3D-07	L25	12	7	6-1-(1)	0
IGKV3D-11	L20	11	7	2-1-(U)	0
IGKV3D-20	A11	12	7	6-1-(1)	2
IGKV4-1	B3	17	7	3-1-(1)	**83**
IGKV5-2	B2	11	7	2-1-(1)	1
IGKV6-21	A10, A26	11	7	2-1-(1)	6
IGKV6D-41	A14	11	7	2-1-(1)	0

[1] Adapted from Tomlinson *et al.* EMBO J., 1995, 14: 4628, incorporated by reference in its entirety. The number in parenthesis refers to canonical structures in CDRL3, if one assuming the most common length (see Example 5 for further detail about CDRL3).

[2] Estimated from sets of human VK sequences compiled from the NCBI database; full set of GI numbers provided in Appendix A.

The 14 most commonly occurring IGKV germline genes (bolded in column 6 of Table 9) account for just over 90% of the usage of the entire repertoire in peripheral blood. From the analysis of Table 9, ten IGKV germline genes were selected for representation as chassis in the currently exemplified library (Table 10). All but V1-12 and V1-27 are among the top 10 most commonly occurring. IGKV germline genes VH2-30, which was tenth in terms of occurrence in peripheral blood, was not included in the currently exemplified embodiment of the library, in order to maintain the proportion of chassis with short (*i.e.,* 11 or 12 residues in length) CDRL1 sequences at about 80% in the final set of 10 chassis. V1-12 was included in its place. V1-17 was more similar to other members of the V1 family that were already selected ; therefore, V1-27 was included, instead of V1-17. In other embodiments, the library could include 12 chassis (*e.g.,* the ten of Table 10 plus V1-17 and V2-30), or a different set of any "N" chassis, chosen strictly by occurrence (Table 9) or any other criteria. The ten chosen VK chassis account for about 80% of the usage in the data set believed to be representative of the entire kappa light chain repertoire.

Table 10. VK Chassis Selected for Use in the Exemplary Library

Chassis	**CDR-L1 Length**	**CDR-L2 Length**	**Canonical Structures**	**Estimated Relative Occurrence in Peripheral Blood**
VK1-5	11	7	2-1-(U)	69
VK1-12	11	7	2-1-(1)	32
VK1-27	11	7	2-1-(1)	27
VK1-33	11	7	2-1-(1)	43
VK1-39	11	7	2-1-(1)	147
VK2-28	16	7	4-1-(1)	62

VK3-11	11	7	2-1-(1)	87
VK3-15	11	7	2-1-(1)	53
VK3-20	12	7	6-1-(1)	195
VK4-1	17	7	3-1-(1)	83

The amino acid sequences of the selected VK chassis enumerated in Table 10 are provided in Table 11.

Table 11. Amino Acid Sequences for VK Chassis Selected for Inclusion in the Exemplary Library

Chassis	FRM1	CDRL1	FRM2	CDRL2	FRM3	CDRL3[1]	SEQ ID NO:
VK1-5	DIQMTQSPSTLSASVGDRVTITC	RASQSISSWLA	WYQQKPGKAPKLLIY	DASSLES	GVPSRFSGSGSGTEFTLTISSLQPDDFATYYC	QYNSYS	
VK1-12	DIQMTQSPSSVSASVGDRVTITC	RASQGISSWLA	WYQQKPGKAPKLLIY	AASSLQS	GVPSRFSGSGSGTDFTLTISSLQPEDFATYYC	QANSFP	
VK1-27	DIQMTQSPSSLSASVGDRVTITC	RASQGISNYLA	WYQQKPGKVPKLLIY	AASTLQS	GVPSRFSGSGSGTDFTLTISSLQPEDVATYYC	KYNSAP	
VK1-33	DIQMTQSPSSLSASVGDRVTITC	QASQDISNYLN	WYQQKPGKAPKLLIY	DASNLET	GVPSRFSGSGSGTDFTFTISSLQPEDIATYYC	QYDNLP	
VK1-39	DIQMTQSPSSLSASVGDRVTITC	RASQSISSYLN	WYQQKPGKAPKLLIY	AASSLQS	GVPSRFSGSGSGTDFTLTISSLQPEDFATYYC	QSYSTP	
VK2-28	DIVMTQSPLSLPVTPGEPASISC	RSSQSLLHSNGYNYLD	WYLQKPGQSPQLLIY	LGSNRAS	GVPDRFSGSGSGTDFTLKISRVEAEDVGVYYC	QALQTP	
VK3-11	EIVLTQSPATLSLSPGERATLSC	RASQSVSSYLA	WYQQKPGQAPRLLIY	DASNRAT	GIPARFSGSGSGTDFTLTISSLEPEDFAVYYC	QRSNWP	
VK3-15	EIVMTQSPATLSVSPGERATLSC	RASQSVSSNLA	WYQQKPGQAPRLLIY	GASTRAT	GIPARFSGSGSGTEFTLTISSLQSEDFAVYYC	QYNNWP	
VK3-20	EIVLTQSPGTLSLSPGERATL	RASQSVSSSYLA	WYQQKPGQAPRLLIY	GASSRAT	GIPDRFSGSGSGTDFTLTISRLEPEDFAVYYC	QYGSSP	

	SC						
VK4-1	DIVMTQSPDSLAVSLGERATINC	KSSQSVLYSSNNKNYLA	WYQQKPGQPPKLLIY	WASTRES	GVPDRFSGSGSGTDFTLTISSLQAEDVAVYYC	QYYSTP	

[1] Note that the portion of the IGKV gene contributing to VKCDR3 is not considered part of the chassis as described herein. The VK chassis is defined as Kabat residues 1 to 88 of the IGKV-encoded sequence, or from the start of FRM1 to the end of FRM3. The portion of the VKCDR3 sequence contributed by the IGKV gene is referred to herein as the L3-VK region.

Example 4: Design of an Exemplary Vλ Chassis Library

This example, describes the design of an exemplary Vλ chassis library. As was previously demonstrated in Examples 1-3, for the VH and VK chassis sequences, the sequence characteristics and occurrence of human IgλV germline-derived sequences in peripheral blood were analyzed. As with the assignment of other sequences set forth herein to germline families, assignment of Vλ sequences to a germline family was performed via SoDA and VBASE2 (Volpe and Kepler, Bioinformatics, 2006, 22: 438; Mollova *et al.*, BMS Systems Biology, 2007, 1S: P30, each incorporated by reference in its entirety). The data are presented in Table 12.

Table 12. IGλV Gene Characteristics and Occurrence in Peripheral Blood

IGλV Gene	Alternative Name	Canonical Structures[1]	Contribution of IGVλ Gene to CDRL3	Estimated Relative Occurrence in Peripheral Blood[2]
IGλV3-1	3R	11-7(*)	8	11.5
IGλV3-21	3H	11-7(*)	9	10.5
IGλV2-14	2A2	14-7(A)	9	10.1
IGλV1-40	1E	14-7(A)	9	7.7
IGλV3-19	3L	11-7(*)	9	7.6
IGλV1-51	1B	13-7(A)	9	7.4
IGλV1-44	1C	13-7(A)	9	7.0
IGλV6-57	6A	13-7(B)	7	6.1
IGλV2-8	2C	14-7(A)	9	4.7
IGλV3-25	3M	11-7(*)	9	4.6
IGλV2-23	2B2	14-7(A)	9	4.3
IGλV3-10	3P	11-7(*)	9	3.4
IGλV4-69	4B	12-11(*)	7	3.0
IGλV1-47	1G	13-7(A)	9	2.9
IGλV2-11	2E	14-7(A)	9	1.3
IGλV7-43	7A	14-7(B)	8	1.3
IGλV7-46	7B	14-7(B)	8	1.1
IGλV5-45	5C	14-11(*)	8	1.0
IGλV4-60	4A	12-11(*)	7	0.7

IGλV10-54	8A	14-7(B)	8	0.7
IGλV8-61	10A	13-7(C)	9	0.7
IGλV3-9	3J	11-7(*)	8	0.6
IGλV1-36	1A	13-7(A)	9	0.4
IGλV2-18	2D	14-7(A)	9	0.3
IGλV3-16	3A	11-7(*)	9	0.2
IGλV3-27		11-7(*)	7	0.2
IGλV4-3	5A	14-11(*)	8	0.2
IGλV5-39	4C	12-11(*)	12	0.2
IGλV9-49	9A	12-12(*)	12	0.2
IGλV3-12	3I	11-7(*)	9	0.1

[1] Adapted from Williams et al. J. Mol. Biol. 1996: 264, 220-32. The (*) indicates that the canonical structure is entirely defined by the lengths of CDRs L1 and L2. When distinct structures are possible for identical L1 and L2 length combinations, the structure present in a given gene is set forth as A, B, or C.

[2] Estimated from a set of human Vλ sequences compiled from the NCBI database; full set of GI codes set forth in Appendix B.

To choose a subset of the sequences from Table 12 to serve as chassis, those represented at less than 1% in peripheral blood (as extrapolated from analysis of published sequences corresponding to the GI codes provided in Appendix B) were first discarded. From the remaining 18 germline sequences, the top occurring genes for each unique canonical structure and contribution to CDRL3, as well as any germline gene represented at more than the 5% level, were chosen to constitute the exemplary Vλ chassis. The list of 11 such sequences is given in Table 13, below. These 11 sequences represent approximately 73% of the repertoire in the examined data set (Appendix B).

Table 13. Vλ Chassis Selected for Use in the Exemplary Library

Chassis	CDRL1 Length	CDRL2 Length	Canonical Structure	Relative Occurrence
Vλ3-1	11	7	11-7(*)	11.5
Vλ3-21	11	7	11-7(*)	10.5
Vλ2-14	14	7	14-7(A)	10.1
Vλ1-40	14	7	14-7(A)	7.7
Vλ3-19	11	7	11-7(*)	7.6
Vλ1-51	13	7	13-7(A)	7.4
Vλ1-44	13	7	13-7(A)	7.0
Vλ6-57	13	7	13-7(B)	6.1
Vλ4-69	12	11	12-11(*)	3.0
Vλ7-43	14	7	14-7(B)	1.3
Vλ5-45	11	11	14-11(*)	1.0

The amino acid sequences of the selected Vλ chassis enumerated in Table 13 are provided in Table 14, below.

Table 14. Amino Acid Sequences for Vλ Chassis Selected for Inclusion in the Exemplary Library

Chassis	FRM1	CDRL1	FRM2	CDRL2	FRM3	CDRL3[2]
Vλ1-40	QSVLTQPPSVSGAPGQRVTISC	TGSSSNIGAGYD---VH	WYQQLPGTAPKLLIY	GN----SNRPS	GVPDRFSGSKSG--TSASLAITGLQAEDEADYYC	QSYDSSLSG
Vλ1-44	QSVLTQPPSASGTPGQRVTISC	SGSSSNIGSNT----VN	WYQQLPGTAPKLLIY	SN----NQRPS	GVPDRFSGSKSG--TSASLAISGLQSEDEADYYC	AAWDDSLNG
Vλ1-51	QSVLTQPPSVSAAPGQKVTISC	SGSSSNIGNNY----VS	WYQQLPGTAPKLLIY	DN----NKRPS	GIPDRFSGSKSG--TSATLGITGLQTGDEADYYC	GTWDSSLSA
Vλ2-14	QSALTQPASVSGSPGQSITISC	TGTSSDVGGYNY---VS	WYQQHPGKAPKLMIY	EV----SNRPS	GVSNRFSGSKSG--NTASLTISGLQAEDEADYYC	SSYTSSSTL
Vλ3-1[1]	SYELTQPPSVSVSPGQTASITC	SGDKLGDKY------A**S**	WYQQKPGQSPVLVIY	QD----SKRPS	GIPERFSGSNSG--NTATLTISGTQAMDEADYYC	QAWDSSTA-
Vλ3-19	SSELTQDPAVSVALGQTVRITC	QGDSLRSYY------AS	WYQQKPGQAPVLVIY	GK----NNRPS	GIPDRFSGSSSG--NTASLTITGAQAEDEADYYC	NSRDSSGNH
Vλ3-21	SYVLTQPPSVSVAPGKTARITC	GGNNIGSKS------VH	WYQQKPGQAPVLVIY	YD----SDRPS	GIPERFSGSNSG--NTATLTISRVEAGDEADYYC	QVWDSSSDH
Vλ4-69	QLVLTQSPSASASLGASVKLTC	TLSSGHSSYA-----IA	WHQQQPEKGPRYLMK	LNSDGSHSKGD	GIPDRFSGSSSG--AERYLTISSLQSEDEADYYC	QTWGTGI--
Vλ6-57	NFMLTQPHSVSESPGKTVTISC	TRSSGSIASNY----VQ	WYQQRPGSSPTTVIY	ED----NQRPS	GVPDRFSGSIDSSSNSASLTISGLKTEDEADYYC	QSYDSSN--
Vλ5-45	QAVLTQPASLSASPGASASLTC	TLRSGINVGTYR---IY	WYQQKPGSPPQYLLR	YKSDSDKQQGS	GVPSRFSGSKDASANAGILLISGLQSEDEADYYC	MIWHSSAS-
Vλ7-43	QTVVTQEPSLTVSPGGTVTLTC	ASSTGAVTSGYY---PN	WFQQKPGQAPRALIY	ST----SNKHS	WTPARFSGSLLG--GKAALTLSGVQPEDEAEYYC	LLYYGGAQ-

[1] The last amino acid in CDRL1 of the Vλ3-1 chassis, S, differs from the corresponding one in the IGλV3-1 germline gene, C. This was done to avoid having a potentially unpaired CYS (C) amino acid in the resulting synthetic light chain.

[2] Note that, as for the VK chassis, the portion of the IGλV gene contributing to VλCDR3 is not considered part of the chassis as described herein. The Vλ chassis is defined as Kabat residues 1 to 88 of the IGλV-encoded sequence, or from the start of FRM1 to the end of FRM3. The portion of the VλCDR3 sequence contributed by the IGλV gene is referred to herein as the L3-Vλ region.

Example 5: Design of a CDRH3 Library

This example describes the design of a CDHR3 library from its individual components. In nature, the CDRH3 sequence is derived from a complex process involving recombination of three different genes, termed IGHV, IGHD and IGHJ. In addition to recombination, these genes may also undergo progressive nucleotide deletions: from the 3' end of the IGHV gene, either end of the IGHD gene, and/or the 5' end of the IGHJ gene. Non-templated nucleotide additions may also occur at the junctions between the V, D and J sequences. Non-templated additions at the V-D junction are referred to as "N1", and those at the D-J junction are referred to as "N2". The D gene segments may be read in three forward and, in some cases, three reverse reading frames.

In the design of the present exemplary library, the codon (nucleotide triplet) or single amino acid was designated as a fundamental unit, to maintain all sequences in the desired reading frame. Thus, all deletions or additions to the gene segments are carried out via the addition or deletion of amino acids or codons, and not single nucleotides. According to the CDRH3 numbering system of this application, CDRH3 extends from amino acid number 95 (when present; see Example 1) to amino acid 102.

Example 5.1: Selection of the DH Segments

In this illustrative example, selection of DH gene segments for use in the library was performed according to principles similar to those used for the selection of the chassis sequences. First, an analysis of IGHD gene usage was performed, using data from Lee *et al.,* Immunogenetics, 2006, 57: 917; Corbett *et al.,* PNAS, 1982, 79: 4118; and Souto-Carneiro *et al.,* J. Immunol., 2004, 172: 6790 (each incorporated by reference in its entirety), with preference for representation in the library given to those IGHD genes most frequently observed in human sequences. Second, the degree of deletion on either end of the IGHD gene segments was estimated by comparison with known heavy chain sequences, using the SoDA algorithm (Volpe *et al.,* Bioinformatics, 2006, 22: 438, incorporated by reference in its entirety) and sequence alignments. For the presently exemplified library, progressively deleted DH segments, as short as three amino acids, were included. As enumerated in the Detailed Description, other embodiments of the invention comprise DH segments with deletions to a different length, for example, about 1, 2, 4, 5, 6, 7, 8, 9, or 10 amino acids. Table 15 shows the relative occurrence of IGHD gene usage in human antibody heavy chain sequences isolated mainly from peripheral blood B cells (list adapted from Lee *et al.,* Immunogenetics, 2006, 57: 917, incorporated by reference in its entirety).

Table 15. Usage of IGHD Genes Based on Relative Occurrence in Peripheral Blood*

IGHD Gene	**Estimated Relative Occurrence in Peripheral Blood[3]**
IGHD3-10	117
IGHD3-22	111
IGHD6-19	95
IGHD6-13	93
IGHD3-3	82
IGHD2-2	63
IGHD4-17	61
IGHD1-26	51
IGHD5-5 / 5-18[1]	49
IGHD2-15	47
IGHD6-6	38
IGHD3-9	32
IGHD5-12	29
IGHD5-24	29
IGHD2-21	28
IGHD3-16	18
IGHD4-23	13
IGHD1-1	9
IGHD1-7	9
IGHD4-4/4-11[2]	7
IGHD1-20	6
IGHD7-27	6
IGHD2-8	4
IGHD6-25	3

[1] Although distinct genes in the genome, the nucleotide sequences of IGHD5-5 and IGHD5-18 are 100% identical and thus indistinguishable in rearranged VH sequences.

[2] IGHD4-4 and IGHD4-11 are also 100% identical.

[3] Adapted from Lee *et al.* Immunogenetics, 2006, 57: 917, by merging the information for distinct alleles of the same IGHD gene.

* IGHD1-14 may also be included in the libraries of the invention.

The translations of the ten most commonly expressed IGHD gene sequences found in naturally occurring human antibodies, in three reading frames, are shown in Table 16. Those reading frames which occur most commonly in peripheral blood have been highlighted in gray. As in Table 15, data regarding IGHD sequence usage and reading frame statistics were derived from Lee *et al.,* 2006, and data regarding IGHD sequence reading frame usage were further complemented by data derived from Corbett *et al.,* PNAS, 1982, 79: 4118 and Souto-Carneiro *et al.,* J. Immunol, 2004, 172: 6790, each of which is incorporated by reference in its entirety.

Table 16. Translations of the Ten Most Common Naturally Occurring IGHD Sequences, in Three Reading Frames (RF)

IGHD	RF 1	SEQ ID NO	RF 2	SEQ ID NO	RF 3	SEQ ID NO
IGHD3-10	VLLWFGELL		YYYGSGSYYN		ITMVRGVII	
IGHD3-22	VLLL###WLLL		YYYDSSGYYY		ITMIVVVIT	
IGHD6-19	GYSSGWY		GIAVAG		V#QWLV	
IGHD6-13	GYSSSWY		GIAAAG		V#QQLV	
IGHD3-03	VLRFLEWLLY		YYDFWSGYYT		ITIFGVVII	
IGHD2-02	WIL##YQLLC		GYCSSTSCYT		DIVVVPAAM	
IGHD4-17	#LR#L		DYGDY		TTVT	
IGHD1-26	GIVGATT		V#WELL		YSGSYY	
IGHD5-5/5-18	VDTAMVT		WIQLWL		GYSYGY	
IGHD2-15	RIL#WW#LLL		GYCSGGSCYS		DIVVVVAAT	

represents a stop codon.
Reading frames highlighted in gray correspond to the most commonly used reading frames.

In the presently exemplified library, the top 10 IGHD genes most frequently used in heavy chain sequences occurring in peripheral blood were chosen for representation in the library. Other embodiments of the library could readily utilize more or fewer D genes. The amino acid sequences of the selected IGHD genes, including the most commonly used reading frames and the total number of variants after progressive N- and C-terminal deletion to a minimum of three residues, are listed in Table 17. As depicted in Table 17, only the most commonly occurring alleles of certain IGHD genes were included in the illustrative library. This is, however, not required, and other embodiments of the invention may utilize IGHD reading frames that occur less frequently in the peripheral blood.

Table 17. D Genes Selected for use in the Exemplary Library

IGHD Gene[1]	Amino Acid Sequence	SEQ ID NO:	Total Number of Variants[2]

IGHD1-26_1	GIVGATT		15
IGHD1-26_3	YSGSYY		10
IGHD2-2_2	GYCSSTSCYT		9[3]
IGHD2-2_3	DIVVVPAAM		28
IGHD2-15_2	GYCSGGSCYS		9
IGHD3-3_3	ITIFGVVII		28
IGHD3-10_1	VLLWFGELL		28
IGHD3-10_2	YYYGSGSYYN		36
IGHD3-10_3	ITMVRGVII		28
IGHD3-22_2	YYYDSSGYYY		36
IGHD4-17_2	DYGDY		6
IGHD5-5_3	GYSYGY		10
IGHD6-13_1	GYSSSWY		15
IGHD6-13_2	GIAAAG		10
IGHD6-19_1	GYSSGWY		15
IGHD6-19_2	GIAVAG		10

[1] The reading frame (RF) is specified as _RF after the name of the gene.

[2] In most cases the total number of variants is given by (N-1) times (N-2) divided by two, where N is the total length in amino acids of the intact D segment.

[3] As detailed herein, the number of variants for segments containing a putative disulfide bond (two C or Cys residues) is limited in this illustrative embodiment.

For each of the selected sequences of Table 17, variants were generated by systematic deletion from the N- and/or C-termini, until there were three amino acids remaining. For example, for the IGHD4-17_2 above, the full sequence DYGDY (SEQ ID NO:___) may be used to generate the progressive deletion variants: DYGD (SEQ ID NO:___), YGDY (SEQ ID NO:___), DYG (SEQ ID NO:___), GDY (SEQ ID NO:___) and YGD (SEQ ID NO:___). In general, for any full-length sequence of size N, there will be a total of (N-1)*(N-2)/2 total variants, including the original full sequence. For the disulfide-loop-encoding segments, as exemplified by reading frame 2 of both IGHD2-2 and IGHD2-15, (*i.e.,* IGHD2-2_2 and IGH2-15_2), the progressive deletions were limited, so as to leave the loop intact *i.e.*, only amino acids N-terminal to the first Cys, or C-terminal to the second Cys, were deleted in the respective DH segment variants. The foregoing strategy was used to avoid the presence of unpaired cysteine residues in the exemplified version of the library. However, as discussed in the Detailed Description, other embodiments of the library may include unpaired cysteine residues, or the substitution of these cysteine residues with other amino acids. In the cases where the truncation of the IGHD gene is limited by the presence of the Cys residues, only 9 variants (including the original full sequence) were generated; *e.g.,* for IGHD2-2_2, the variants would be: GYCSSTSCYT (SEQ ID NO:___), GYCSSTSCY (SEQ ID NO:___), YCSSTSCYT (SEQ ID NO:___), CSSTSCYT (SEQ ID NO:___),

GYCSSTSC (SEQ ID NO:___), YCSSTSCY (SEQ ID NO:___), CSSTSCY (SEQ ID NO:___), YCSSTSC (SEQ ID NO:___) and CSSTSC (SEQ ID NO:___).

According to the criteria outlined above, 293 DH sequences were obtained from the selected IGHD gene segments, including the original IGHD gene segments. Certain sequences are redundant. For example, it is possible to obtain the YYY variant from either IGHD3-10_2 (full sequence **YYY**GSGSYYN (SEQ ID NO:___)), or in two different ways from IGHD3-22_2 (SEQ ID NO:___) (**YYY**DSSG**YYY**). When redundant sequences are removed, the number of unique DH segment sequences in this illustrative embodiment of the library is 278. These sequences are enumerated in Table 18.

Table 18. DH Gene Segments Used in the Presently Exemplified Library*

DH Segment Designation[1]	Peptide	SEQ ID NO:	DH Segment Designation	Peptide	SEQ ID NO:
IGHD1-26_1-1	ATT		IGHD3-10_2-20	YYGSG	
IGHD1-26_1-2	GAT		IGHD3-10_2-21	YYYGS	
IGHD1-26_1-3	GIV		IGHD3-10_2-22	GSGSYY	
IGHD1-26_1-4	IVG		IGHD3-10_2-23	SGSYYN	
IGHD1-26_1-5	VGA		IGHD3-10_2-24	YGSGSY	
IGHD1-26_1-6	GATT		IGHD3-10_2-25	YYGSGS	
IGHD1-26_1-7	GIVG		IGHD3-10_2-26	YYYGSG	
IGHD1-26_1-8	IVGA		IGHD3-10_2-27	GSGSYYN	
IGHD1-26_1-9	VGAT		IGHD3-10_2-28	YGSGSYY	
IGHD1-26_1-10	GIVGA		IGHD3-10_2-29	YYGSGSY	
IGHD1-26_1-11	IVGAT		IGHD3-10_2-30	YYYGSGS	
IGHD1-26_1-12	VGATT		IGHD3-10_2-31	YGSGSYYN	
IGHD1-26_1-13	GIVGAT		IGHD3-10_2-32	YYGSGSYY	
IGHD1-26_1-14	IVGATT		IGHD3-10_2-33	YYYGSGSY	
IGHD1-26_1-15	GIVGATT		IGHD3-10_2-34	YYGSGSYYN	
IGHD1-26_3-1	YSG		IGHD3-10_2-35	YYYGSGSYY	
IGHD1-26_3-2	YSGS		IGHD3-10_2-36	YYYGSGSYYN	

IGHD1-26_3-3	YSGSY		IGHD3-10_3-1	GVI	
IGHD1-26_3-4	YSGSYY		IGHD3-10_3-2	ITM	
IGHD2-02_2-1	CSSTSC		IGHD3-10_3-3	MVR	
IGHD2-02_2-2	CSSTSCY		IGHD3-10_3-4	RGV	
IGHD2-02_2-3	YCSSTSC		IGHD3-10_3-5	TMV	
IGHD2-02_2-4	CSSTSCYT		IGHD3-10_3-6	VII	
IGHD2-02_2-5	GYCSSTSC		IGHD3-10_3-7	VRG	
IGHD2-02_2-6	YCSSTSCY		IGHD3-10_3-8	GVII	
IGHD2-02_2-7	GYCSSTSCY		IGHD3-10_3-9	ITMV	
IGHD2-02_2-8	YCSSTSCYT		IGHD3-10_3-10	MVRG	
IGHD2-02_2-9	GYCSSTSCYT		IGHD3-10_3-11	RGVI	
IGHD2-02_3-1	AAM		IGHD3-10_3-12	TMVR	
IGHD2-02_3-2	DIV		IGHD3-10_3-13	VRGV	
IGHD2-02_3-3	IVV		IGHD3-10_3-14	ITMVR	
IGHD2-02_3-4	PAA		IGHD3-10_3-15	MVRGV	
IGHD2-02_3-5	VPA		IGHD3-10_3-16	RGVII	
IGHD2-02_3-6	VVP		IGHD3-10_3-17	TMVRG	
IGHD2-02_3-7	VVV		IGHD3-10_3-18	VRGVI	
IGHD2-02_3-8	DIVV		IGHD3-10_3-19	ITMVRG	
IGHD2-02_3-9	IVVV		IGHD3-10_3-20	MVRGVI	
IGHD2-02_3-10	PAAM		IGHD3-10_3-21	TMVRGV	
IGHD2-02_3-11	VPAA		IGHD3-10_3-22	VRGVII	
IGHD2-02_3-12	VVPA		IGHD3-10_3-23	ITMVRGV	
IGHD2-02_3-13	VVVP		IGHD3-10_3-24	MVRGVII	
IGHD2-02_3-14	DIVVV		IGHD3-10_3-25	TMVRGVI	
IGHD2-02_3-15	IVVVP		IGHD3-10_3-26	ITMVRGVI	
IGHD2-02_3-16	VPAAM		IGHD3-10_3-27	TMVRGVII	
IGHD2-02_3-17	VVPAA		IGHD3-10_3-28	ITMVRGVII	
IGHD2-02_3-18	VVVPA		IGHD3-22_2-1	DSS	
IGHD2-02_3-19	DIVVVP		IGHD3-22_2-2	GYY	
IGHD2-02_3-20	IVVVPA		IGHD3-22_2-3	SGY	
IGHD2-02_3-21	VVPAAM		IGHD3-22_2-4	SSG	
IGHD2-02_3-22	VVVPAA		IGHD3-22_2-5	YDS	
IGHD2-02_3-23	DIVVVPA		IGHD3-22_2-6	YYD	
IGHD2-02_3-24	IVVVPAA		IGHD3-22_2-7	DSSG	
IGHD2-02_3-25	VVVPAAM		IGHD3-22_2-8	GYYY	
IGHD2-02_3-26	DIVVVPAA		IGHD3-22_2-9	SGYY	

IGHD2-02_3-27	IVVVPAAM		IGHD3-22_2-10	SSGY	
IGHD2-02_3-28	DIVVVPAAM		IGHD3-22_2-11	YDSS	
IGHD2-15_2-1	CSGGSC		IGHD3-22_2-12	YYDS	
IGHD2-15_2-2	CSGGSCY		IGHD3-22_2-13	YYYD	
IGHD2-15_2-3	YCSGGSC		IGHD3-22_2-14	DSSGY	
IGHD2-15_2-4	CSGGSCYS		IGHD3-22_2-15	SGYYY	
IGHD2-15_2-5	GYCSGGSC		IGHD3-22_2-16	SSGYY	
IGHD2-15_2-6	YCSGGSCY		IGHD3-22_2-17	YDSSG	
IGHD2-15_2-7	GYCSGGSCY		IGHD3-22_2-18	YYDSS	
IGHD2-15_2-8	YCSGGSCYS		IGHD3-22_2-19	YYYDS	
IGHD2-15_2-9	GYCSGGSCYS		IGHD3-22_2-20	DSSGYY	
IGHD3-03_3-1	FGV		IGHD3-22_2-21	SSGYYY	
IGHD3-03_3-2	GVV		IGHD3-22_2-22	YDSSGY	
IGHD3-03_3-3	IFG		IGHD3-22_2-23	YYDSSG	
IGHD3-03_3-4	ITI		IGHD3-22_2-24	YYYDSS	
IGHD3-03_3-5	TIF		IGHD3-22_2-25	DSSGYYY	
IGHD3-03_3-6	VVI		IGHD3-22_2-26	YDSSGYY	
IGHD3-03_3-7	FGVV		IGHD3-22_2-27	YYDSSGY	
IGHD3-03_3-8	GVVI		IGHD3-22_2-28	YYYDSSG	
IGHD3-03_3-9	IFGV		IGHD3-22_2-29	YDSSGYYY	
IGHD3-03_3-10	ITIF		IGHD3-22_2-30	YYDSSGYY	
IGHD3-03_3-11	TIFG		IGHD3-22_2-31	YYYDSSGY	
IGHD3-03_3-12	VVII		IGHD3-22_2-32	YYDSSGYYY	
IGHD3-03_3-13	FGVVI		IGHD3-22_2-33	YYYDSSGYY	
IGHD3-03_3-14	GVVII		IGHD3-22_2-34	YYYDSSGYYY	
IGHD3-03_3-15	IFGVV		IGHD4-17_2-1	DYG	
IGHD3-03_3-16	ITIFG		IGHD4-17_2-2	GDY	
IGHD3-03_3-17	TIFGV		IGHD4-17_2-3	YGD	
IGHD3-03_3-18	FGVVII		IGHD4-17_2-4	DYGD	
IGHD3-03_3-19	IFGVVI		IGHD4-17_2-5	YGDY	
IGHD3-03_3-20	ITIFGV		IGHD4-17_2-6	DYGDY	
IGHD3-03_3-21	TIFGVV		IGHD5-5_3-1	SYG	
IGHD3-03_3-22	IFGVVII		IGHD5-5_3-2	YGY	

IGHD3-03_3-23	ITIFGVV		IGHD5-5_3-3	YSY	
IGHD3-03_3-24	TIFGVVI		IGHD5-5_3-4	GYSY	
IGHD3-03_3-25	ITIFGVVI		IGHD5-5_3-5	SYGY	
IGHD3-03_3-26	TIFGVVII		IGHD5-5_3-6	YSYG	
IGHD3-03_3-27	ITIFGVVII		IGHD5-5_3-7	GYSYG	
IGHD3-10_1-1	ELL		IGHD5-5_3-8	YSYGY	
IGHD3-10_1-2	FGE		IGHD5-5_3-9	GYSYGY	
IGHD3-10_1-3	GEL		IGHD6-13_1-1	SSS	
IGHD3-10_1-4	LLW		IGHD6-13_1-2	SSW	
IGHD3-10_1-5	LWF		IGHD6-13_1-3	SWY	
IGHD3-10_1-6	VLL		IGHD6-13_1-4	SSSW	
IGHD3-10_1-7	WFG		IGHD6-13_1-5	SSWY	
IGHD3-10_1-8	FGEL		IGHD6-13_1-6	YSSS	
IGHD3-10_1-9	GELL		IGHD6-13_1-7	GYSSS	
IGHD3-10_1-10	LLWF		IGHD6-13_1-8	SSSWY	
IGHD3-10_1-11	LWFG		IGHD6-13_1-9	YSSSW	
IGHD3-10_1-12	VLLW		IGHD6-13_1-10	GYSSSW	
IGHD3-10_1-13	WFGE		IGHD6-13_1-11	YSSSWY	
IGHD3-10_1-14	FGELL		IGHD6-13_1-12	GYSSSWY	
IGHD3-10_1-15	LLWFG		IGHD6-19_1-1	GWY	
IGHD3-10_1-16	LWFGE		IGHD6-19_1-2	GYS	
IGHD3-10_1-17	VLLWF		IGHD6-19_1-3	SGW	
IGHD3-10_1-18	WFGEL		IGHD6-19_1-4	YSS	
IGHD3-10_1-19	LLWFGE		IGHD6-19_1-5	GYSS	
IGHD3-10_1-20	LWFGEL		IGHD6-19_1-6	SGWY	
IGHD3-10_1-21	VLLWFG		IGHD6-19_1-7	SSGW	
IGHD3-10_1-22	WFGELL		IGHD6-19_1-8	YSSG	
IGHD3-10_1-23	LLWFGEL		IGHD6-19_1-9	GYSSG	
IGHD3-10_1-24	LWFGELL		IGHD6-19_1-10	SSGWY	
IGHD3-10_1-25	VLLWFGE		IGHD6-19_1-11	YSSGW	
IGHD3-10_1-26	LLWFGELL		IGHD6-19_1-12	GYSSGW	
IGHD3-10_1-27	VLLWFGEL		IGHD6-19_1-13	YSSGWY	
IGHD3-10_1-28	VLLWFGELL		IGHD6-19_1-14	GYSSGWY	
IGHD3-10_2-1	GSG		IGHD6-19_2-1	AVA	
IGHD3-10_2-2	GSY		IGHD6-19_2-2	GIA	
IGHD3-10_2-3	SGS		IGHD6-19_2-3	IAV	
IGHD3-10_2-4	SYY		IGHD6-19_2-4	VAG	
IGHD3-10_2-5	YGS		IGHD6-19_2-5	AVAG	
IGHD3-10_2-6	YYG		IGHD6-19_2-6	GIAV	
IGHD3-10_2-7	YYN		IGHD6-19_2-7	IAVA	
IGHD3-10_2-8	YYY		IGHD6-19_2-8	GIAVA	
IGHD3-10_2-9	GSGS		IGHD6-19_2-9	IAVAG	
IGHD3-10_2-10	GSYY		IGHD6-19_2-10	GIAVAG	
IGHD3-10_2-11	SGSY		IGHD6-13_2-1	AAA	
IGHD3-10_2-12	SYYN		IGHD6-13_2-2	AAG	

IGHD3-10_2-13	YGSG		IGHD6-13_2-3	IAA	
IGHD3-10_2-14	YYGS		IGHD6-13_2-4	AAAG	
IGHD3-10_2-15	YYYG		IGHD6-13_2-5	GIAA	
IGHD3-10_2-16	GSGSY		IGHD6-13_2-6	IAAA	
IGHD3-10_2-17	GSYYN		IGHD6-13_2-7	GIAAA	
IGHD3-10_2-18	SGSYY		IGHD6-13_2-8	IAAAG	
IGHD3-10_2-19	YGSGS		IGHD6-13_2-9	GIAAAG	

[1] The sequence designation is formatted as follows: (IGHD Gene Name)_(Reading Frame)-(Variant Number)

* Note that the origin of certain variants is rendered somewhat arbitrary when redundant segments are deleted from the library (*i.e.,* certain segments may have their origins with more than one parent, including the one specified in the table).

Table 19 shows the length distribution of the 278 DH segments selected according to the methods described above.

Table 19 Length Distributions of DH Segments Selected for Inclusion in the Exemplary Library

DH Size	Number of Occurrences
3	78
4	64
5	50
6	38
7	27
8	20
9	12
10	4

As specified above, based on the CDRH3 numbering system defined in this application, IGHD-derived amino acids (*i.e.,* DH segments) are numbered beginning with position 97, followed by positions 97A, 97B, etc. In the currently exemplified embodiment of the library, the shortest DH segment has three amino acids: 97, 97A and 97B, while the longest DH segment has 10 amino acids: 97, 97A, 97B, 97C, 97D, 97E, 97F, 97G, 97H and 97I.

Example 5.2: Selection of the H3-JH Segments

There are six human germline IGHJ genes. During *in vivo* assembly of antibody genes, these segments are progressively deleted at their 5’ end. In this exemplary embodiment of the library, IGHJ gene segments with no deletions, or with 1, 2, 3, 4, 5, 6, or 7 deletions (at the amino acid level), yielding JH segments as short as 13 amino acids, were included (Table 20). Other embodiments of the invention, in which the IGHJ gene segments are progressively deleted (at their 5' / N-terminal end) to yield 15, 14, 12, or 11 amino acids are also contemplated.

Table 20. IGHJ Gene Segments Selected for use in the Exemplary Library

IGHJ Segment	[H3-JH]-[FRM4][1]	SEQ ID NO:	H3-JH	SEQ ID NO:
JH1 parent or JH1_1	AEYFQHWGQGTLVTVSS		AEYFQH	
JH1_2	EYFQHWGQGTLVTVSS		EYFQH	
JH1_3	YFQHWGQGTLVTVSS		YFQH	
JH1_4	FQHWGQGTLVTVSS		FQH	
JH1_5	QHWGQGTLVTVSS		QH	
JH2 parent or JH2_1	YWYFDLWGRGTLVTVSS		YWYFDL	
JH2_2	WYFDLWGRGTLVTVSS		WYFDL	
JH2_3	YFDLWGRGTLVTVSS		YFDL	
JH2_4	FDLWGRGTLVTVSS		FDL	
JH2_5	DLWGRGTLVTVSS		DL	
JH3 parent or JH3_1	AFDVWGQGTMVTVSS		AFDV	
JH3_2	FDVWGQGTMVTVSS		FDV	
JH3_3	DVWGQGTMVTVSS		DV	
JH4 parent or JH4_1	YFDYWGQGTLVTVSS		YFDY	
JH4_2	FDYWGQGTLVTVSS		FDY	
JH4_3	DYWGQGTLVTVSS		DY	
JH5 parent or JH5_1	NWFDSWGQGTLVTVSS		NWFDS	
JH5_2	WFDSWGQGTLVTVSS		WFDS	
JH5_3	FDSWGQGTLVTVSS		FDS	
JH5_4	DSWGQGTLVTVSS		DS	
JH6 parent or JH6_1	YYYYYGMDVWGQGTTVTVSS		YYYYYGMDV	
JH6_2	YYYYGMDVWGQGTTVTVSS		YYYYGMDV	
JH6_3	YYYGMDVWGQGTTVTVSS		YYYGMDV	
JH6_4	YYGMDVWGQGTTVTVSS		YYGMDV	
JH6_5	YGMDVWGQGTTVTVSS		YGMDV	
JH6_6	GMDVWGQGTTVTVSS		GMDV	
JH6_7	MDVWGQGTTVTVSS		MDV	
JH6_8	DVWGQGTTVTVSS		DV	

[1] H3-JH is defined as the portion of the IGHJ segment included within the Kabat definition of CDRH3; FRM4 is defined as the portion of the IGHJ segment encoding framework region four.

According to the CDRH3 numbering system of this application, the contribution of, for example, JH6_1 to CDRH3, would be designated by positions 99F, 99E, 99D, 99C, 99B, 99A, 100, 101 and 102 (Y, Y, Y, Y, Y, G, M, D and V, respectively). Similarly, the JH4_3 sequence would contribute amino acid positions 101 and 102 (D and Y, respectively) to CDRH3. However, in all cases of the exemplified library, the JH segment will contribute amino acids 103 to 113 to the FRM4 region, in accordance with the standard Kabat numbering system for antibody variable regions (Kabat, op. cit. 1991). This may not be the case in other embodiments of the library.

Example 5.3: Selection of the N1 and N2 Segments

While the consideration of V-D-J recombination enhanced by mimicry of the naturally occurring process of progressive deletion (as exemplified above) can generate enormous diversity, the diversity of the CDRH3 sequences *in vivo* is further amplified by non-templated addition of a varying number of nucleotides at the V-D junction and the D-J junction.

N1 and N2 segments located at the V-D and D-J junctions, respectively, were identified in a sample containing about 2,700 antibody sequences (Jackson *et al.*, J. Immunol. Methods, 2007, 324: 26) also analyzed by the SoDA method of Volpe *et al.*, Bioinformatics, 2006, 22: 438-44; (both Jackson *et al.,* and Volpe *et al.,* are incorporated by reference in their entireties). Examination of these sequences revealed patterns in the length and composition of N1 and N2. For the construction of the currently exemplified CDRH3 library, specific short amino acid sequences were derived from the above analysis and used to generate a number of N1 and N2 segments that were incorporated into the CDRH3 design, using the synthetic scheme described herein.

As described in the Detailed Description, certain embodiments of the invention include N1 and N2 segments with rationally designed length and composition, informed by statistical biases in these parameters that are found by comparing naturally occurring N1 and N2 segments in human antibodies. According to data compiled from human databases (see, *e.g.,* Jackson *et al.,* J. Immunol Methods, 2007, 324: 26, incorporated by reference in its entirety), there are an average of about 3.02 amino acid insertions for N1 and about 2.4 amino acid insertions for N2, not taking into account insertions of two nucleotides or less. Figure 2 shows the length distributions of the N1 and N2 regions in human antibodies. In this exemplary embodiment of the invention, N1 and N2 were fixed to a length of 0, 1, 2, or 3 amino acids. The naturally occurring composition of these sequences in human antibodies was used as a guide for the inclusion of different amino acid residues.

The naturally occurring composition of single amino acid, two amino acids, and three amino acids N1 additions is defined in Table 21, and the naturally occurring composition of the corresponding N2 additions is defined in Table 22. The most frequently occurring duplets in the N1 and N2 set are compiled in Table 23.

Table 21. Composition of Naturally Occurring 1, 2, and 3 Amino Acid N1 Additions*

Position 1	Number of Occurrences	Position 2	Number of Occurrences	Position 3	Number of Occurrences
R	251	**G**	97	**G**	101
G	249	**P**	67	**R**	66
P	173	**R**	67	**P**	47
L	130	**S**	42	**S**	47
S	117	**L**	39	**L**	38
A	84	**V**	33	**A**	33
V	62	**E**	24	**V**	28
K	61	**A**	21	**T**	27
I	55	**D**	18	**E**	24
Q	51	**I**	18	**D**	22
T	51	**T**	18	**K**	18
D	50	**K**	16	**F**	14
E	49	**Y**	16	**I**	13
F	3	**H**	13	**W**	13
H	32	**F**	12	**N**	10
N	30	**Q**	11	**Y**	10
W	28	**N**	5	**H**	8
Y	21	**W**	5	**Q**	5
M	16	**C**	4	**C**	3
C	3	**M**	4	**M**	3
	1546		530		530

* Defined as the sequence C-terminal to "CARX", or equivalent, of VH, wherein "X" is the "tail" (*e.g.*, D, E, G, or no amino acid residue).

Table 22. Composition of Naturally Occurring 1, 2, and 3 Amino Acid N2 Additions*

Position 1	Number of Occurrences	Position 2	Number of Occurrences	Position 3	Number of Occurrences
G	242	G	244	G	156
P	219	P	138	P	79
R	180	R	86	S	54
L	132	S	85	R	51
S	123	T	77	L	49
A	97	L	74	A	41
T	78	A	69	T	31
V	75	V	46	V	29
E	57	E	41	D	23
D	56	Y	38	E	23
F	54	D	36	W	23
H	54	K	30	Q	19
Q	53	F	29	F	17
I	49	W	27	Y	17
N	45	H	24	H	16
Y	40	I	23	I	11
K	35	Q	23	K	11
W	29	N	21	N	8
M	20	M	8	C	6
C	6	C	5	M	6
	1644		1124		670

* Defined as the sequence C-terminal to the D segment but not encoded by IGHJ genes.

Table 23. Top Twenty-Five Naturally Occurring N1 and N2 Duplets

Sequence	Number of Occurrences	Cumulative Frequency	Individual Frequency
GG	17	0.037	0.037
PG	15	0.070	0.033
RG	15	0.103	0.033
PP	13	0.132	0.029
GP	12	0.158	0.026
GL	11	0.182	0.024
PT	10	0.204	0.022
TG	10	0.226	0.022
GV	9	0.246	0.020
RR	9	0.266	0.020
SG	8	0.284	0.018
RP	7	0.299	0.015
IG	6	0.312	0.013
GS	6	0.325	0.013
SR	6	0.338	0.013
PA	6	0.352	0.013
LP	6	0.365	0.013
VG	6	0.378	0.013
KG	6	0.389	0.011
GW	5	0.400	0.011
FP	5	0.411	0.011
LG	5	0.422	0.011
RS	5	0.433	0.011
TP	5	0.444	0.011
EG	5	0.455	0.011

Example 5.3.1 Selection of the N1 Segments

Analysis of the identified N1 segments, located at the junction between V and D, revealed that the eight most frequently occurring amino acid residues were G, R, S, P, L, A, T and V (Table 21). The number of amino acid additions in the N1 segment was frequently none, one, two, or three (Figure 2). The addition of four or more amino acids was relatively rare. Therefore, in the currently exemplified embodiment of the library, the N1 segments were designed to include zero, one, two or three amino acids. However, in other embodiments, N1 segments of four, five, or more amino acids may also be utilized. G and P were always among the most commonly occurring amino acid residues in the N1 regions. Thus, in the present exemplary embodiment of the library, the N1 segments that are dipeptides are of the form GX, XG, PX, or XP, where X is any of the eight most commonly occurring amino acids listed above. Due to the fact that G residues were observed more frequently than P residues, the tripeptide members of the exemplary N1 library have the form GXG, GGX, or XGG, where X is, again, one of the eight most frequently occurring amino acid residues listed above. The resulting set of N1 sequences used in the present exemplary embodiment of the library, include the "zero" addition amounts to 59 sequences, which are listed in Table 24.

Table 24. N1 Sequences Selected for Inclusion in the Exemplary Library

Segment Type	Sequences	Number
"Zero"	(no addition) V segment joins directly to D segment	1
Monomers	G, P, R, A, S, L, T, V	8
Dimers	GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP	28
Trimers	GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV	22

In accordance with the CDRH3 numbering system of the application, the sequences enumerated in Table 24 contribute the following positions to CDRH3: the monomers contribute position 96, the dimers to 96 and 96A, and the trimers to 96, 96A and 96B. In alternative embodiments, where tetramers and longer segments could be included among the N1 sequences, the corresponding numbers would go on to include 96C, and so on.

Example 5.3.2 Selection of the N2 Segments

Similarly, analysis of the identified N2 segments, located at the junction between D and J, revealed that the eight most frequently occurring amino acid residues were also G, R, S, P, L, A, T and V (Table 22). The number of amino acid additions in the N2 segment was also frequently none, one, two, or three (Figure 2). For the design of the N2 segments in the exemplary library, an expanded set of sequences was utilized. Specifically, the sequences in Table 25 were used, in addition to the 59 sequences enumerated in Table 24, for N1.

Table 25. Extra Sequences in N2 Additions

Segment Type	Sequence	Number New	Number Total
Monomers	D, E, F, H, I, K, M, Q, W, Y	10	18

Dimers	AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS	54	82
Trimers	AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT	18	40

The presently exemplified embodiment of the library, therefore, contains 141 total N2 sequences, including the "zero" state. One of ordinary skill in the art will readily recognize that these 141 sequences may also be used in the N1 region, and that such embodiments are within the scope of the invention. In addition, the length and compositional diversity of the N1 and N2 sequences can be further increased by utilizing amino acids that occur less frequently than G, R, S, P, L, A, T and V, in the N1 and N2 regions of naturally occurring antibodies, and including N1 and N2 segments of four, five, or more amino acids in the library. Tables 21 to 23 and Figure 2 provides information about the composition and length of the N1 and N2 sequences in naturally occurring antibodies that is useful for the design of additional N1 and N2 regions which mimic the natural composition and length.

In accordance with the CDRH3 numbering system of the application, N2 sequences will begin at position 98 (when present) and extend to 98A (dimers) and 98B (trimers). Alternative embodiments may occupy positions 98C, 98D, and so on.

Example 5.4. A CDRH3 Library

When the "tail" (*i.e.,* G/D/E/-) is considered, the CDRH3 in the exemplified library may be represented by the general formula:

[G/D/E/-]-[N1]-[DH]-[N2]-[H3-JH]

In the currently exemplified, non-limiting, embodiment of the library, [G/D/E/-] represents each of the four possible terminal amino acid "tails"; N1 can be any of the 59 sequences in Table 24; DH can be any of the 278 sequences in Table 18; N2 can be any of the 141 sequences in Tables 24 and 25; and H3-JH can be any of the 28 H3-JH sequences in Table 20. The total theoretical diversity or repertoire size of this CDRH3 library is obtained by multiplying the variations at each of the components, *i.e.,* 4 × 59 × 278 × 141 × 28 = 2.59 × 10^8.

However, as described in the previous examples, redundancies may be eliminated from the library. In the presently exemplified embodiment, the tail and N1 segments were combined, and redundancies were removed from the library. For example, considering the VH chassis, tail, and N1 regions, the sequence [VH_Chassis]-[G] may be obtained in two different ways: [VH_Chassis] + [G] + [nothing] or [VH_Chassis] + [nothing] + [G]. Removal of redundant sequences resulted in a total of 212 unique [G/D/E/-]-[N1] segments out of the 236 possible combinations (*i.e.,* 4 tails × 59 N1). Therefore, the actual diversity of the presently exemplified CDRH3 library is 212 × 278 × 141 × 28 = 2.11 × 10^8. Figure 23 depicts the frequency of occurrence of different CDRH3 lengths in this library, versus the preimmune repertoire of Lee *et al.*

Table 26 further illustrates specific exemplary sequences from the CDRH3 library described above, using the CDRH3 numbering system of the present application. In instances where a position is not used, the hyphen symbol (-) is included in the table instead.

Table 26. Examples of Designed CDRH3 Sequences According to the Library Exemplified in Examples 1 to 5

	[Tail]	[N1]			[DH]										[N2]			[H3-JH]									CDRH3 Length
	95	96	96A	96B	97	97A	97B	97C	97D	97E	97F	97G	97H	97I	98	98A	98B	99E	99D	99C	99B	99A	99	100	101	102	
No. 1	G	-	-	-	Y	Y	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	D	V	6
No. 2	D	G	-	-	G	Y	C	S	G	G	S	C	Y	S	Y	-	-	-	-	-	-	-		F	Q	H	16
No. 3	E	R	-	-	I	T	I	F	G	V	-	-	-	-	G	G	-	-	-	-	-	-	Y	F	D	Y	14
No. 4	-	P	P	-	V	L	L	W	F	G	E	L	L	-	D	-	-	-	-	-	-	-	-	-	D	L	14
No. 5	G	G	S	G	Y	Y	Y	G	S	G	S	Y	Y	N	P	-	-	-	-	-	A	E	Y	F	Q	H	21
No. 6	D	-	-	-	R	G	V	I	I	-	-	-	-	-	M	-	-	Y	Y	Y	Y	Y	G	M	D	V	16
No. 7	E	S	G	-	Y	Y	Y	D	S	S	G	Y	Y	Y	T	G	L	-	-	-	-	W	Y	F	D	L	21
No. 8	-	S	-	-	D	Y	G	D	Y	-	-	-	-	-	S	I	-	-	-	-	-	-	-	F	D	I	11
No. 9	-	P	G	-	W	F	G	-	-	-	-	-	-	-	P	S	-	-	-	-	Y	Y	G	M	D	V	13
No. 10	-	-	-	-	C	S	G	G	S	C	-	-	-	-	A	Y	-	-	-	-	-	N	W	F	D	P	13

Example 6: Design of VKCDR3 Libraries

This example describes the design of a number of exemplary VKCDR3 libraries. As specified in the Detailed Description, the actual version(s) of the VKCDR3 library made or used in particular embodiments of the invention will depend on the objectives for the use of the library. In this example the Kabat numbering system for light chain variable regions was used.

In order to facilitate examination of patterns of occurrence, human kappa light chain sequences were obtained from the publicly available NCBI database (Appendix A). As for the heavy chain sequences (Example 2), each of the sequences obtained from the publicly available database was assigned to its closest germline gene, on the basis of sequence identity. The amino acid compositions at each position were then determined within each kappa light chain subset.

Example 6.1.: A Minimalist VKCDR3 Library

This example describes the design of a “minimalist” VKCDR3 library, wherein the VKCDR3 repertoire is restricted to a length of nine residues. Examination of the VKCDR3 lengths of human sequences shows that a dominant proportion (over 70%) has nine amino acids within the Kabat definition of CDRL3: positions 89 through 97. Thus, the currently exemplified minimalist design considers only VKCDR3 of length nine. Examination of human kappa light chain sequences shows that there are not strong biases in the usage of IGKJ genes; there are five such IKJ genes in humans. Table 27 depicts IGKJ gene usage amongst three data sets, namely Juul *et al.* (Clin. Exp. Immunol., 1997, 109: 194, incorporated by reference in its entirety), Klein and Zachau (Eur. J. Immunol., 1993, 23: 3248, incorporated by reference in its entirety), and the kappa light chain data set provided in Appendix A (labeled LUA).

Table 27. IGKJ Gene Usage in Various Data Sets

Gene	Klein	Juul	LUA
IGKJ1	35.0%	29.0%	29.3%
IGKJ2	25.0%	23.0%	24.1%
IGKJ3	7.0%	8.0%	12.1%
IGKJ4	26.0%	24.0%	26.5%
IGKJ5	6.0%	18.0%	8.0%

Thus, a simple combinatorial of "M" VK chassis and the 5 IGKJ genes would generate a library of size M × 5. In the Kabat numbering system, for VKCDR3 of length nine, amino acid number 96 is the first encoded by the IGKJ gene. Examination of the amino acid occupying this position in human sequences showed that the seven most common residues are L, Y, R, W, F, P, and I, cumulatively accounting for about 85% of the residues found in position 96. The remaining 13 amino acids account for the other 15%. The occurrence of all 20 amino acids at position 96 is presented in Table 28.

Table 28. Occurrence of 20 Amino Acid Residues at Position 96 in Human VK Data Set

Type	Number	Percent	Cumulative
L	**333**	**22.3**	**22.3**
Y	**235**	**15.8**	**38.1**
R	**222**	**14.9**	**52.9**
W	**157**	**10.5**	**63.5**
F	**148**	**9.9**	**73.4**
I	**96**	**6.4**	**79.8**
P	90	6.0	85.9
Q	53	3.6	89.4
N	39	2.6	92.0
H	31	2.1	94.1
V	21	1.4	95.5
G	20	1.3	96.8
C	14	0.9	97.8
K	7	0.5	98.3
S	6	0.4	98.7
A	5	0.3	99.0
D	5	0.3	99.3
E	5	0.3	99.7
T	5	0.3	100.0
M	0	0.0	100.0

To determine the origins of the seven residues most commonly found in position 96, known human IGKJ amino acid sequences were examined (Table 29).

Table 29. Known Human IGKJ Amino Acid Sequences

Gene	Sequence
IGKJ1	WTFGQGTKVEIK
IGKJ2	YTFGQGTKLEIK
IGKJ3	FTFGPGTKVDIK
IGKJ4	LTFGGGTKVEIK
IGKJ5	ITFGQGTRLEIK

Without being bound by theory, five of the seven most commonly occurring amino acids found in position 96 of rearranged human sequences appear to originate from the first amino acid encoded by each of the five human IGKJ genes, namely, W, Y, F, L, and I.

Less evident were the origins of the P and R residues. Without being bound by theory, most of the human IGKV gene nucleotide sequences end with the sequence CC, which occurs after (*i.e.,* 3' to) the end of the last full codon (*e.g.,* that encodes the C-terminal residue shown in Table 11). Therefore, regardless of which nucleotide is placed after this sequence (*i.e.,* CCX, where X may be any nucleotide) the codon will encode a proline (P) residue. Thus, when the IGKJ gene undergoes progressive deletion (just as in the IGHJ of the heavy chain; see Example 5), the first full amino acid is lost and, if no deletions have occurred in the IGKV gene, a P residue will result.

To determine the origin of the arginine residue at position 96, the origin of IGKJ genes in rearranged kappa light chain sequences containing R at position 96 were analyzed. The analysis indicated that R occurred most frequently at position 96 when the IGKJ gene was IGKJ1. The germline W (position 1; Table 29) for IGKJ1 is encoded by **T**GG. Without being bound by theory, a single nucleotide change of **T** to C (yielding CGG) or A (yielding AGG) will, therefore, result in codons encoding Arg (R). A change to G (yielding GGG) results in a codon encoding Gly (G). R occurs about ten times more often at position 96 in human sequences than G (when the IGKJ gene is IGKJ1), and it is encoded by CGG more often than AGG. Therefore, without being bound by theory, C may originate from one of the aforementioned two Cs at the end of IGKV gene. However, regardless of the mechanism(s) of occurrence, R and P are among the most frequently observed amino acid types at position 96, when the length of VKCDR3 is 9. Therefore, a minimalist VKCDR3 library may be represented by the following amino acid sequence:

[VK_Chassis]-[L3-VK]-[F/L/I/R/W/Y/P]-[TFGGGTKVEIK]

In this sequence, VK_Chassis represents any selected VK chassis (for non-limiting examples, see Table 11), specifically Kabat residues 1 to 88 encoded by the IGKV gene. L3-VK represents the portion of the VKCDR3 encoded by the chosen IGKV gene (in this embodiment, residues 89-95). F/L/I/R/W/Y/P represents any one of amino residues F, L, I, R, W, Y, or P. In this exemplary representation, IKJ4 (minus the first residue) has been depicted. Without being bound by theory, apart from IGKJ4 being among the most commonly used IGKJ genes in humans, the GGG amino acid sequence is expected to lead to larger conformational flexibility than any of the alternative IGKJ genes, which contain a GXG amino acid sequence, where X is an amino acid other than G. In some embodiments, it may be advantageous to produce a minimalist pre-immune repertoire with a higher degree of conformational flexibility. Considering the ten VK chassis depicted in Table 11, one implementation of the minimalist VKCDR3 library would have 70 members resulting from the combination of 10 VK chassis by 7 junction (position 96) options and one IGKJ-derived sequence (*e.g.,* IGKJ4). Although this embodiment of the library has been depicted using IGKJ4, it is possible to design a minimalist VKCDR3 library using one of the other four IGKJ sequences. For example, another embodiment of the library may have 350 members (10 VK chassis by 7 junctions by 5 IGKJ genes).

One of ordinary skill in the art will readily recognize that one or more minimalist VKCDR3 libraries may be constructed using any of the IGKJ genes. Using the notation above, these minimalist VKCDR3 libraries may have sequences represented by, for example:

***JK1*:** **[VK_Chassis]-[L3-VK]-[F/L/I/R/W/Y/P]-[TFGQGTKVEIK]**;
JK2: **[VK_Chassis]- [L3-VK]-[F/L/I/R/W/Y/P]-[TFGQGTKLEIK]**;
JK3: **[VK_Chassis]- [L3-VK]-[F/L/I/R/W/Y/P]-[TFGPGTKVDIK]**; and
JK5: **[VK_Chassis]- [L3-VK]-[F/L/I/R/W/Y/P]-[TFGQGTRLEIK]**.

Example 6.2: A VKCDR3 Library of About 10^5 Complexity

In this example, the nine residue VKCDR3 repertoire described in Example 6.1 is expanded to include VKCDR3 lengths of eight and ten residues. Moreover, while the previously enumerated VKCDR3 library included the VK chassis and portions of the IGKJ gene not contributing to VKCDR3, the presently exemplified version focuses only on residues comprising a portion of VKCDR3. This embodiment may be favored, for example, when recombination with a vector which already contains VK chassis sequences and constant region sequences is desired.

While the dominant length of VKCDR3 sequences in humans is nine amino acids, other lengths appear at measurable rates that cumulatively approach almost 30% of kappa light chain sequences. In particular, VKCDR3 of lengths 8 and 10 represent, respectively, about 8.5% and about 16% of sequences in representative samples (Figure 3). Thus, a more complex VKCDR3 library includes CDR lengths of 8 to 10 amino acids; this library accounts for over 95% of the length distribution observed in typical collections of human VKCDR3 sequences. This library also enables the inclusion of additional variation outside of the junction between the VK and JK genes. The present example describes such a library. The library comprises 10 sub-libraries, each designed around one of the 10 exemplary VK chassis depicted in Table 11. Clearly, the approach exemplified here can be generalized to consider M different chassis, where M may be less than or more than 10.

To characterize the variability within the polypeptide segment occupying Kabat positions 89 to 95, human kappa light chain sequence collections derived from each of the ten germline sequences of Example 3 were aligned and compared separately (*i.e.,* within the germline group). This analysis enabled us to discern the patterns of sequence variation at each individual position in each kappa light chain sequence, grouped by germline. The table below shows the results for sequences derived from IGKV1-39.

Table 30. Percent Occurrence of Amino Acid Types in IGKV1-39-Derived Sequences

Amino Acid	P89	P90	P91	P92	P93	P94	P95
A	0	0	1	0	0	4	1
C	0	0	0	0	0	0	0
D	0	0	1	1	3	0	0
E	0	1	0	0	0	0	0
F	0	0	0	5	0	2	0
G	0	0	2	1	2	0	0
H	1	1	0	4	0	0	0
I	0	0	1	0	4	5	1
K	0	0	0	1	2	0	0
L	3	0	0	1	1	3	7
M	0	0	0	0	0	1	0
N	0	0	3	2	6	2	0
P	0	0	0	0	0	4	85
Q	96	97	0	0	0	0	0
R	0	0	0	0	5	0	2
S	0	0	80	4	65	6	3
T	0	0	9	0	10	65	1
V	0	0	0	0	0	1	1
W	0	0	0	0	0	0	0
Y	0	0	2	80	0	3	0

For example, at position 89, two amino acids, Q and L, account for about 99% of the observed variability, and thus in the currently exemplified library (see below), only

Q and L were included in position 89. In larger libraries, of course, additional, less frequently occurring amino acid types (*e.g.,* H), may also be included.

Similarly, at position 93 there is more variation, with amino acid types S, T, N, R and I being among the most frequently occurring. The currently exemplified library thus aimed to include these five amino acids at position 93, although clearly others could be included in more diverse libraries. However, because this library was constructed via standard chemical oligonucleotide synthesis, one is bound by the limits of the genetic code, so that the actual amino acid set represented at position 93 of the exemplified library consists of S, T, N, R, P and H, with P and H replacing I (see exemplary 9 residue VKCDR3 in Table 32, below). This limitation may be overcome by using codon-based synthesis of oligonucleotides, as described in Example 6.3, below. A similar approach was followed at the other positions and for the other sequences: analysis of occurrences of amino acid type per position, choice from among most frequently occurring subset, followed by adjustment as dictated by the genetic code.

As indicated above, the library employs a practical and facile synthesis approach using standard oligonucleotide synthesis instrumentation and degenerate oligonucleotides. To facilitate description of the library, the IUPAC code for degenerate nucleotides, as given in Table 31, will be used.

Table 31. Degenerate Base Symbol Definition

IUPAC Symbol	Base Pair Composition
A	A (100%)
C	C (100%)
G	G (100%)
T	T (100%)
R	A (50%) G (50%)
Y	C (50%) T (50%)
W	A (50%) T (50%)
S	C (50%) G (50%)
M	A (50%) C (50%)
K	G (50%) T (50%)
B	C (33%) G (33%) T (33%) (*)
D	A (33%) G (33%) T (33%)
H	A (33%) C (33%) T (33%)
V	A (33%) C (33%) G (33%)
N	A (25%) C (25%) G (25%) T (25%)

(*) 33% is short hand here for 1/3 (*i.e.,* 33.3333 … %)

Using the VK1-39 chassis with VKCDR3 of length nine as an example, the VKCDR3 library may be represented by the following four oligonucleotides (left column in Table 32), with the corresponding amino acids encoded at each position of CDRL3 (Kabat numbering) provided in the columns on the right.

Table 32. Exemplary Oligonucleotides Encoding a VK1-39 CDR3 Library

Oligonucleotide Sequence	89	90	91	92	93	94	95	95A	96	97
CWGSAAWCATHCMVTABTCCTTWCACT	LQ	EQ	ST	FSY	HNPRST	IST	P	-	FY	T
CWGSAAWCATHCMVTABTCCTMTCACT	LQ	EQ	ST	FSY	HNPRST	IST	P	-	IL	T
CWGSAAWCATHCMVTABTCCTWGGACT	LQ	EQ	ST	FSY	HNPRST	IST	P	-	WR	T
CWGSAAWCATHCMVTABTCCTCBTACT	LQ	EQ	ST	FSY	HNPRST	IST	P	PLR	-	T

For example, the first codon (CWG) of the first nucleotide of Table 32, corresponding to Kabat position 89, represents 50% CTG and 50% CAG, which encode Leu (L) and Gln (Q), respectively. Thus, the expressed polypeptide would be expected to have L and Q each about 50% of the time. Similarly, for Kabat position 95A of the fourth oligonucleotide, the codon CBT represents 1/3 each of CCT, CGT and CTT, corresponding in turn to 1/3 each of Pro (P), Leu (L) and Arg (R) upon translation. By multiplying the number of options available at each position of the peptide sequence, one can obtain the complexity, in peptide space, contributed by each oligonucleotide. For the VK1-39 example above, the numbers are 864 for the first three oligonucleotides and 1,296 for the fourth oligonucleotide. Thus, the oligonucleotides encoding VK1-39 CDR3s of length nine contribute 3,888 members to the library. However, as shown in Table 32, sequences with L or R at position 95A (when position 96 is empty) are identical to those with L or R at position 96 (and 95A empty). Therefore, the 3,888 number overestimates the LR contribution and the actual number of unique members is slightly lower, at 3,024. As depicted in Table 33, for the complete list of oligonucleotides that represent VKCDR3 of sizes 8, 9, and 10, for all 10 VK chassis, the overall complexity is about 1.3×10^5 or 1.2×10^5 unique sequences after correcting for over-counting of the LR contribution for the size 9 VKCDR3.

Table 33. Degenerate Oligonucleotides Encoding an Exemplary VKCDR3 Library

Chassis	CDR L3 Length	Junction Type (1)	Degenerate Oligonucleotide	SEQ ID NO:	89	90	91	92	93	94	95	95A	96	97
VK1-5	8	1	CASCASTMCVRTRSTTWCTWCACT		HQ	HQ	SY	DGHNRS	AGST	FY	-	-	FY	T
VK1-5	8	2	CASCASTMCVRTRSTTWCMTCACT		HQ	HQ	SY	DGHNRS	AGST	FY	-	-	IL	T
VK1-5	8	3	CASCASTMCVRTRSTTWCWGGACT		HQ	HQ	SY	DGHNRS	AGST	FY	-	-	WR	T
VK1-5	8	4	CASCASTMCVRTRSTTWCYCTACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	-	-	T
VK1-5	9	1	CASCASTMCVRTRSTTWCYCTTWCACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	-	FY	T
VK1-5	9	2	CASCASTMCVRTRSTTWCYCTMTCACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	-	IL	T
VK1-5	9	3	CASCASTMCVRTRSTTWCYCTWGGACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	-	WR	T
VK1-5	9	4	CASCASTMCVRTRSTTWCYCTYCTACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	PS	-	T
VK1-5	10	1	CASCASTMCVRTRSTTWCYCTCBTTWCACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	PLR	FY	T
VK1-5	10	2	CASCASTMCVRTRSTTWCYCTCBTMTCACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	PLR	IL	T
VK1-5	10	3	CASCASTMCVRTRSTTWCYCTCBTWGGACT		HQ	HQ	SY	DGHNRS	AGST	FY	PS	PLR	WR	T

VK1-12	8	1	CASCASDCTRVCARTTTSTWCACT		HQ	HQ	AST	ADGNST	NS	FL	-	-	FY	T
VK1-12	8	2	CASCASDCTRVCARTTTSMTCACT		HQ	HQ	AST	ADGNST	NS	FL	-	-	IL	T
VK1-12	8	3	CASCASDCTRVCARTTTSWGGACT		HQ	HQ	AST	ADGNST	NS	FL	-	-	WR	T
VK1-12	8	4	CASCASDCTRVCARTTTSCCTACT		HQ	HQ	AST	ADGNST	NS	FL	P	-	-	T
VK1-12	9	1	CASCASDCTRVCARTTTSCCTTWCACT		HQ	HQ	AST	ADGNST	NS	FL	P	-	FY	T
VK1-12	9	2	CASCASDCTRVCARTTTSCCTMTCACT		HQ	HQ	AST	ADGNST	NS	FL	P	-	IL	T
VK1-12	9	3	CASCASDCTRVCARTTTSCCTWGGACT		HQ	HQ	AST	ADGNST	NS	FL	P	-	WR	T
VK1-12	9	4	CASCASDCTRVCARTTTSCCTCBTACT		HQ	HQ	AST	ADGNST	NS	FL	P	PLR	-	T
VK1-12	10	1	CASCASDCTRVCARTTTSCCTCBTTWCACT		HQ	HQ	AST	ADGNST	NS	FL	P	PLR	FY	T
VK1-12	10	2	CASCASDCTRVCARTTTSCCTCBTMTCACT		HQ	HQ	AST	ADGNST	NS	FL	P	PLR	IL	T
VK1-12	10	3	CASCASDCTRVCARTTTSCCTCBTWGGACT		HQ	HQ	AST	ADGNST	NS	FL	P	PLR	WR	T
VK1-27	8	1	CASMAGTWCRRTASKGBATWCACT		HQ	KQ	FY	DGNS	RST	AGV	-	-	FY	T
VK1-27	8	2	CASMAGTWCRRTASKGBAMTCACT		HQ	KQ	FY	DGNS	RST	AGV	-	-	IL	T
VK1-27	8	3	CASMAGTWCRRTASKGBAWGGACT		HQ	KQ	FY	DGNS	RST	AGV	-	-	WR	T
VK1-27	8	4	CASMAGTWCRRTASKGBACCTACT		HQ	KQ	FY	DGNS	RST	AGV	P	-	-	T
VK1-27	9	1	CASMAGTWCRRTASKGBACCTTWCACT		HQ	KQ	FY	DGNS	RST	AGV	P	-	FY	T
VK1-27	9	2	CASMAGTWCRRTASKG		HQ	KQ	FY	DGNS	RST	AGV	P	-	IL	T

			BACCTMTCACT											
VK1-27	9	3	CASMAGTWCRRTASKGBACCTWGGACT		HQ	KQ	FY	DGNS	RST	AGV	P	-	WR	T
VK1-27	9	4	CASMAGTWCRRTASKGBACCTCBTACT		HQ	KQ	FY	DGNS	RST	AGV	P	PLR	-	T
VK1-27	10	1	CASMAGTWCRRTASKGBACCTCBTTWCACT		HQ	KQ	FY	DGNS	RST	AGV	P	PLR	FY	T
VK1-27	10	2	CASMAGTWCRRTASKGBACCTCBTMTCACT		HQ	KQ	FY	DGNS	RST	AGV	P	PLR	IL	T
VK1-27	10	3	CASMAGTWCRRTASKGBACCTCBTWGGACT		HQ	KQ	FY	DGNS	RST	AGV	P	PLR	WR	T
VK1-33	8	1	CASCWTTMCRATRVCBWTTWCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	-	-	FY	T
VK1-33	8	2	CASCWTTMCRATRVCBWTMTCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	-	-	IL	T
VK1-33	8	3	CASCWTTMCRATRVCBWTWGGACT		HQ	HL	SY	DN	ADGNST	DFHLVY	-	-	WR	T
VK1-33	8	4	CASCWTTMCRATRVCBWTCCTACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	-	-	T
VK1-33	9	1	CASCWTTMCRATRVCBWTCCTTWCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	-	FY	T
VK1-33	9	2	CASCWTTMCRATRVCBWTCCTMTCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	-	IL	T
VK1-33	9	3	CASCWTTMCRATRVCBWTCCTWGGACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	-	WR	T
VK1-33	9	4	CASCWTTMCRATRVCBWTCCTCBTACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	PLR	-	T
VK1-33	10	1	CASCWTTMCRATRVCBWTCCTCBTTWCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	PLR	FY	T
VK1-33	10	2	CASCWTTMCRATRVCBWTCCTCBTMTCACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	PLR	IL	T
VK1-33	10	3	CASCWTTMCRATRVCBWTCCTCBTWGGACT		HQ	HL	SY	DN	ADGNST	DFHLVY	P	PLR	WR	T

VK1-39	8	1	CWGSAAWCATHCMVTABTTWCACT		LQ	EQ	ST	FSY	HNPRST	IST	-	-	FY	T
VK1-39	8	2	CWGSAAWCATHCMVTABTMTCACT		LQ	EQ	ST	FSY	HNPRST	IST	-	-	IL	T
VK1-39	8	3	CWGSAAWCATHCMVTABTWGGACT		LQ	EQ	ST	FSY	HNPRST	IST	-	-	WR	T
VK1-39	8	4	CWGSAAWCATHCMVTABTCCTACT		LQ	EQ	ST	FSY	HNPRST	IST	P	-	-	T
VK1-39	9	1	CWGSAAWCATHCMVTABTCCTTWCACT		LQ	EQ	ST	FSY	HNPRST	IST	P	-	FY	T
VK1-39	9	2	CWGSAAWCATHCMVTABTCCTMTCACT		LQ	EQ	ST	FSY	HNPRST	IST	P	-	IL	T
VK1-39	9	3	CWGSAAWCATHCMVTABTCCTWGGACT		LQ	EQ	ST	FSY	HNPRST	IST	P	-	WR	T
VK1-39	9	4	CWGSAAWCATHCMVTABTCCTCBTACT		LQ	EQ	ST	FSY	HNPRST	IST	P	PLR	-	T
VK1-39	10	1	CWGSAAWCATHCMVTABTCCTCBTTWCACT		LQ	EQ	ST	FSY	HNPRST	IST	P	PLR	FY	T
VK1-39	10	2	CWGSAAWCATHCMVTABTCCTCBTMTCACT		LQ	EQ	ST	FSY	HNPRST	IST	P	PLR	IL	T
VK1-39	10	3	CWGSAAWCATHCMVTABTCCTCBTWGGACT		LQ	EQ	ST	FSY	HNPRST	IST	P	PLR	WR	T
VK3-11	8	1	CASCASAGWRGKRVCTSGTWCACT		HQ	HQ	RS	GRS	ADGNST	SW	-	-	FY	T
VK3-11	8	2	CASCASAGWRGKRVCTSGMTCACT		HQ	HQ	RS	GRS	ADGNST	SW	-	-	IL	T
VK3-11	8	3	CASCASAGWRGKRVCTSGWGGACT		HQ	HQ	RS	GRS	ADGNST	SW	-	-	WR	T
VK3-11	8	4	CASCASAGWRGKRVCTSGCCTACT		HQ	HQ	RS	GRS	ADGNST	SW	P	-	-	T
VK3-11	9	1	CASCASAGWRGKRVCTSGCCTTWCACT		HQ	HQ	RS	GRS	ADGNST	SW	P	-	FY	T
VK3-11	9	2	CASCASAGWRGKRVCT		HQ	HQ	RS	GRS	ADGNS	SW	P	-	IL	T

			SGCCTMTCACT						T					
VK3-11	9	3	CASCASAGWRGKRVCTSGCCTWGGACT		HQ	HQ	RS	GRS	ADGNST	SW	P	-	WR	T
VK3-11	9	4	CASCASAGWRGKRVCTSGCCTCBTACT		HQ	HQ	RS	GRS	ADGNST	SW	P	PLR	-	T
VK3-11	10	1	CASCASAGWRGKRVCTSGCCTCBTTWCACT		HQ	HQ	RS	GRS	ADGNST	SW	P	PLR	FY	T
VK3-11	10	2	CASCASAGWRGKRVCTSGCCTCBTMTCACT		HQ	HQ	RS	GRS	ADGNST	SW	P	PLR	IL	T
VK3-11	10	3	CASCASAGWRGKRVCTSGCCTCBTWGGACT		HQ	HQ	RS	GRS	ADGNST	SW	P	PLR	WR	T
VK3-15	8	1	CASCASTMCVRTRRKTGGTWCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	-	-	FY	T
VK3-15	8	2	CASCASTMCVRTRRKTGGMTCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	-	-	IL	T
VK3-15	8	3	CASCASTMCVRTRRKTGGWGGACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	-	-	WR	T
VK3-15	8	4	CASCASTMCVRTRRKTGGCCTACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	-	-	T
VK3-15	9	1	CASCASTMCVRTRRKTGGCCTTWCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	-	FY	T
VK3-15	9	2	CASCASTMCVRTRRKTGGCCTMTCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	-	IL	T
VK3-15	9	3	CASCASTMCVRTRRKTGGCCTWGGACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	-	WR	T
VK3-15	9	4	CASCASTMCVRTRRKTGGCCTCBTACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	PLR	-	T
VK3-15	10	1	CASCASTMCVRTRRKTGGCCTCBTTWCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	PLR	FY	T
VK3-15	10	2	CASCASTMCVRTRRKTGGCCTCBTMTCACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	PLR	IL	T
VK3-15	10	3	CASCASTMCVRTRRKTGGCCTCBTWGGACT		HQ	HQ	SY	DGHNRS	DEGKNRS	W	P	PLR	WR	T

VK3-20	8	1	CASCASTWCGRTRVKKCATWCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	-	-	FY	T
VK3-20	8	2	CASCASTWCGRTRVKKCAMTCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	-	-	IL	T
VK3-20	8	3	CASCASTWCGRTRVKKCAWGGACT		HQ	HQ	FY	DG	ADEGKNRST	AS	-	-	WR	T
VK3-20	8	4	CASCASTWCGRTRVKKCACCTACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	-	-	T
VK3-20	9	1	CASCASTWCGRTRVKKCACCTTWCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	-	FY	T
VK3-20	9	2	CASCASTWCGRTRVKKCACCTMTCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	-	IL	T
VK3-20	9	3	CASCASTWCGRTRVKKCACCTWGGACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	-	WR	T
VK3-20	9	4	CASCASTWCGRTRVKKCACCTCBTACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	PLR	-	T
VK3-20	10	1	CASCASTWCGRTRVKKCACCTCBTTWCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	PLR	FY	T
VK3-20	10	2	CASCASTWCGRTRVKKCACCTCBTMTCACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	PLR	IL	T
VK3-20	10	3	CASCASTWCGRTRVKKCACCTCBTWGGACT		HQ	HQ	FY	DG	ADEGKNRST	AS	P	PLR	WR	T
VK2-28	8	1	ATGCASRBTCKTSASABTTWCACT		M	HQ	AGISTV	LR	DEHQ	IST	-	-	FY	T
VK2-28	8	2	ATGCASRBTCKTSASABTMTCACT		M	HQ	AGISTV	LR	DEHQ	IST	-	-	IL	T
VK2-28	8	3	ATGCASRBTCKTSASABTWGGACT		M	HQ	AGISTV	LR	DEHQ	IST	-	-	WR	T
VK2-28	8	4	ATGCASRBTCKTSASABTCCTACT		M	HQ	AGISTV	LR	DEHQ	IST	P	-	-	T
VK2-28	9	1	ATGCASRBTCKTSASABTCCTTWCACT		M	HQ	AGISTV	LR	DEHQ	IST	P	-	FY	T
VK2-28	9	2	ATGCASRBTCKTSASA		M	HQ	AGI	LR	DEHQ	IST	P	-	IL	T

			BTCCTMTCACT				STV							
VK2-28	9	3	ATGCASRBTCKTSASABTCCTWGGACT		M	HQ	AGISTV	LR	DEHQ	IST	P	-	WR	T
VK2-28	9	4	ATGCASRBTCKTSASABTCCTCBTACT		M	HQ	AGISTV	LR	DEHQ	IST	P	PLR	-	T
VK2-28	10	1	ATGCASRBTCKTSASABTCCTCBTTWCACT		M	HQ	AGISTV	LR	DEHQ	IST	P	PLR	FY	T
VK2-28	10	2	ATGCASRBTCKTSASABTCCTCBTMTCACT		M	HQ	AGISTV	LR	DEHQ	IST	P	PLR	IL	T
VK2-28	10	3	ATGCASRBTCKTSASABTCCTCBTWGGACT		M	HQ	AGISTV	LR	DEHQ	IST	P	PLR	WR	T
VK4-1	8	1	CASCASTWCTWCRVCABTTWCACT		HQ	HQ	FY	FY	ADGNST	IST	-	-	FY	T
VK4-1	8	2	CASCASTWCTWCRVCABTMTCACT		HQ	HQ	FY	FY	ADGNST	IST	-	-	IL	T
VK4-1	8	3	CASCASTWCTWCRVCABTWGGACT		HQ	HQ	FY	FY	ADGNST	IST	-	-	WR	T
VK4-1	8	4	CASCASTWCTWCRVCABTCCTACT		HQ	HQ	FY	FY	ADGNST	IST	P	-	-	T
VK4-1	9	1	CASCASTWCTWCRVCABTCCTTWCACT		HQ	HQ	FY	FY	ADGNST	IST	P	-	FY	T
VK4-1	9	2	CASCASTWCTWCRVCABTCCTMTCACT		HQ	HQ	FY	FY	ADGNST	IST	P	-	IL	T
VK4-1	9	3	CASCASTWCTWCRVCABTCCTWGGACT		HQ	HQ	FY	FY	ADGNST	IST	P	-	WR	T
VK4-1	9	4	CASCASTWCTWCRVCABTCCTCBTACT		HQ	HQ	FY	FY	ADGNST	IST	P	PLR	-	T
VK4-1	10	1	CASCASTWCTWCRVCABTCCTCBTTWCACT		HQ	HQ	FY	FY	ADGNST	IST	P	PLR	FY	T
VK4-1	10	2	CASCASTWCTWCRVCABTCCTCBTMTCACT		HQ	HQ	FY	FY	ADGNST	IST	P	PLR	IL	T
VK4-1	10	3	CASCASTWCTWCRVCABTCCTCBTWGGACT		HQ	HQ	FY	FY	ADGNST	IST	P	PLR	WR	T

[Alternative for VK1-33] (2)														
VK1-33	8	1	CASCWATMCRATRVCBWTTWCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	-	-	FY	T
VK1-33	8	2	CASCWATMCRATRVCBWTMTCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	-	-	IL	T
VK1-33	8	3	CASCWATMCRATRVCBWTWGGACT		HQ	QL	SY	DN	ADGNST	DFHLVY	-	-	WR	T
VK1-33	8	4	CASCWATMCRATRVCBWTCCTACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	-	-	T
VK1-33	9	1	CASCWATMCRATRVCBWTCCTTWCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	-	FY	T
VK1-33	9	2	CASCWATMCRATRVCBWTCCTMTCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	-	IL	T
VK1-33	9	3	CASCWATMCRATRVCBWTCCTWGGACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	-	WR	T
VK1-33	9	4	CASCWATMCRATRVCBWTCCTCBTACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	PLR	-	T
VK1-33	10	1	CASCWATMCRATRVCBWTCCTCBTTWCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	PLR	FY	T
VK1-33	10	2	CASCWATMCRATRVCBWTCCTCBTMTCACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	PLR	IL	T
VK1-33	10	3	CASCWATMCRATRVCBWTCCTCBTWGGACT		HQ	QL	SY	DN	ADGNST	DFHLVY	P	PLR	WR	T

(1) Junction type 1 has position 96 as FY, type 2 as IL, type 3 as RW, and type 4 has a deletion.

(2) Two embodiments are shown for the VK1-33 library. In one embodiment, the second codon was CWT. In another embodiment, it was CWA or CWG.

Example 6.3: More Complex VKCDR3 Libraries

This example demonstrates how a more faithful representation of amino acid variation at each position may be obtained by using a codon-based synthesis approach (Virnekas *et al.* Nucleic Acids Res., 1994, 22: 5600). This synthetic scheme also allows for finer control of the proportions of particular amino acids included at a position. For example, as described above for the VK1-39 sequences, position 89 was designed as 50% Q and 50% L; however, as Table 30 shows, Q is used much more frequently than L. The more complex VKCDR3 libraries of the present example account for the different relative occurrence of Q and L, for example, 90% Q and 10% L. Such control is better exercised within codon-based synthetic schemes, especially when multiple amino acid types are considered.

This example also describes an implementation of a codon-based synthetic scheme, using the ten VK chassis described in Table 11. Similar approaches, of course, can be implemented with more or fewer such chassis. As indicated in the Detailed Description, a unique aspect of the design of the present libraries, as well as those of the preceding examples, is the germline or chassis-based aspect, which is meant to preserve more of the integrity and variation of actual human kappa light chain sequences. This is in contrast to other codon-based synthesis or degenerate oligonucleotide synthesis approaches that have been described in the literature and that aim to produce "one-size-fits-all" (*e.g.,* consensus) kappa light chain libraries (*e.g.,*, Knappik, *et al.*, J Mol Biol, 2000, 296: 57; Akamatsu *et al.,* J Immunol, 1993, 151: 4651).

With reference to Table 30, obtained for VK1-39, one can thus design the length nine VKCDR3 library of Table 34. Here, for practical reasons, the proportions at each position are denoted in multiples of five percentage points. As better synthetic schemes are developed, finer resolution may be obtained – for example to resolutions of one, two, three, or four percent.

Table 34. Amino Acid Composition (%) at Each VKCDR3 Position for VK1-39 Library With CDR Length of Nine Residues

Amino Acid	89	90	91	92	93	94	95	96 (*)	97 (*)
A						5			5
D				5	5				
E		5				5			

F				5				10	
G			5	5	5			5	
H	5	5		5		5			
I					5			5	
K					5				
L	10					5	10	20	
M									
N			**0**	**0**	5	**0**		5	
P						5	85	5	
Q	85	90						5	
R					5		5	10	
S			80	5	60	5			5
T			10		10	65			90
V								5	
W								15	
Y			5	75		5		15	
Number Different	**3**	**3**	**4**	**6**	**8**	**8**	**3**	**11**	**3**

(*) The composition of positions 96 and 97, determined largely by junction and IGKJ diversity, could be the same for length 9 VK CDR3 of all chassis.

The library of Table 34 would have 1.37×10^6 unique polypeptide sequences, calculated by multiplying together the numbers in the bottom row of the table.

The underlined **0** entries for Asn (N) at certain positions represent regions where the possibility of having N-linked glycosylation sites in the VKCDR3 has been minimized or eliminated. Peptide sequences with the pattern N-X-(S or T)-Z, where X and Z are different from P, may undergo post-translational modification in a number of expression systems, including yeast and mammalian cells. Moreover, the nature of such modification depends on the specific cell type and, even for a given cell type, on culture conditions. N-linked glycosylation may be disadvantageous when it occurs in a region of the antibody molecule likely to be involved in antigen binding (*e.g.,* a CDR), as the function of the antibody may then be influenced by factors that may be difficult to control. For example, considering position 91 above, one can observe that position 92 is never P. Position 94 is not P in 95% of the cases. However, position 93 is S or T in 75 % (65 + 10) of the cases. Thus, allowing N at position 91 would generate the undesirable motif N-X-(T/S)-Z (with both X and Z distinct from P), and a zero occurrence has therefore been implemented, even though N is observed with some frequency in actual human sequences (see Table 30). A similar argument applies for N at positions 92 and 94. It should be appreciated, however, that if the antibody library were to be expressed in a system incapable of N-linked glycosylation, such as bacteria, or under culture conditions in which N-linked glycosylation did not occur, this consideration may not apply. However, even in the event that the organism used to express libraries with potential N-linked glycosylation sites is incapable of N-linked glycosylation (*e.g.,* bacteria), it may still be desirable to avoid N-X-(S/T) sequences, as the antibodies isolated from such libraries may be expressed in different systems (*e.g.,* yeast, mammalian cells) later (*e.g.,* toward clinical development), and the presence of carbohydrate moieties in the variable domains, and the CDRs in particular, may lead to unwanted modifications of activity. These embodiments are also included within the scope of the invention. To our knowledge, VKCDR3 libraries known in the art have not considered this effect, and thus a proportion of their members may have the undesirable qualities mentioned above.

We also designed additional sub-libraries, related to the library outlined in Table 34, for VKCDR3 of lengths 8 and 10. In these embodiments, the compositions at positions 89 to 94 and 97 remain the same as those depicted in Table 34. Additional diversity, introduced at positions 95 and 95A, the latter being defined for VKCDR3 of length 10 only, are illustrated in Table 35.

Table 35. Amino Acid Composition (%) for VK1-39 Libraries of Lengths 8 and 10

Amino Acid	**Position 95 – Length 8 (*)**		**Position 95 – Length 10 (**)**	**Position 95A – Length 10**
A				
D				
E				
F	5			
G				5
H				
I	10			5
K				
L	20		10	10
M				
N				
P	25		85	60
Q				
R	10		5	10
S	5			5
T				5
V	5			
W	10			
Y	10			
Number Different	**9**		**3**	**8**

(*) Position 96 is deleted in VKCDR3 of size 8.
(**) This is the same composition as in VKCDR3 of size 9.

The total number of unique members in the VK1-39 library of length 8, thus, can be obtained as before, and is 3.73×10^5 (or, $3 \times 3 \times 4 \times 6 \times 8 \times 8 \times 9 \times 3$). Similarly, the complexity of the VK1-39 library of length 10 would be 10.9×10^6 (or 8 times that of the library of size 9, as there is additional 8-fold variation at the insertion position 95A). Thus, there would be a total of 12.7×10^6 unique members in the overall VK1-39 library, as obtained by summing the number of unique members for each of the specified lengths. In certain embodiments of the invention, it may be preferable to create the individual sub-libraries of lengths 8, 9 and 10 separately, and then mix the sub-libraries in proportions that reflect the length distribution of VKCDR3 in human sequences; for example, in ratios approximating the 1:9:2 distribution that occurs in natural VKCDR3 sequences (see Figure 3). The present invention provides the compositions and methods for one of ordinary skill synthesizing VKCDR3 libraries corresponding to other VK chassis.

Example 7: A Minimalist VλCDR3 Library

This example describes the design of a minimalist VλCDR3 library. The principles used in designing this library (or more complex Vλ libraries) are similar to those used to design the VKCDR3 libraries. However, unlike the VK genes, the contribution of the IgλV segment to CDRL3 is not constrained to a fixed number of amino acids. Therefore, length variation may be obtained in a minimalist VλCDR3 library even when only considering combinations between Vλ chassis and Jλ sequences.

Examination of the VλCDR3 lengths of human sequences shows that lengths of 9 to 12 account for almost about 95% of sequences, and lengths of 8 to 12 account for about 97% of sequences (Figure 4). Table 36 shows the usage (percent occurrence) of the six known IGλJ genes in the rearranged human lambda light chain sequences compiled from the NCBI database (see Appendix B), and Table 37 shows the sequences encoded by the genes.

Table 36. IGλJ Gene Usage in the Lambda Light Chain Sequences Compiled from the NCBI Database (see Appendix B)

Gene_Allele	LUA
Jλ1_01	20.2%
Jλ2_01	42.2%
Jλ3_02	36.2%
Jλ6_01	0.6%
Jλ7_01	0.9%

Table 37. Observed Human IGλJ Amino Acid Sequences

Gene	Sequence
IGλJ1-01	**YVFGTGTKVTVL**
IGλJ2-01	**VVFGGGTKLTVL**
IGλJ 3-01	WVFGGGTKLTVL
IGλJ3-02	**VVFGGGTKLTVL**
IGλJ6-01	NVFGSGTKVTVL
IGλJ7-01	AVFGGGTQLTVL
IGλJ7-02	AVFGGGTQLTAL

IGλJ3-01 and IGλJ7-02 are not represented among the sequences that were analyzed; therefore, they were not included in Table 36. As illustrated in Table 36, IGλJ1-01, IGλJ2-01, and IGλJ3-02 are over-represented in their usage, and have thus been bolded in Table 37. In some embodiments of the invention, for example, only these three over-represented sequences may be utilized. In other embodiments of the invention, one may use all six segments, any 1, 2, 3, 4, or 5 of the 6 segments, or any combination thereof may be utilized.

As shown in Table 14, the portion of CDRL3 contributed by the IGλV gene segment is 7, 8, or 9 amino acids. The remainder of CDRL3 and FRM4 are derived from the IGλJ sequences (Table 37). The IGλJ sequences contribute either one or two amino acids to CDRL3. If two amino acids are contributed by IGλJ, the contribution is from the N-terminal two residues of the IGλJ segment: YV (IGλJ1-01), VV (IGλJ2-01), WV (IGλJ3-01), VV (IGλJ3-02), or AV (IGλJ7-01 and IGλJ7-02). If one amino acid is contributed from IGλJ, it is a V residue, which is formed after the deletion of the N-terminal residue of a IGλJ segment.

In this non-limiting exemplary embodiment of the invention, the FRM4 segment was fixed as FGGGTKLTVL, corresponding to IGλJ2-01 and IGλJ3-02.

Seven of the 11 selected chassis (Vλ1-40, Vλ3-19, Vλ3-21, Vλ6-57, Vλ1-44, Vλ1-51, and Vλ4-69) have an additional two nucleotides following the last full codon. In four of those seven cases, analysis of the data set provided in Appendix B showed that the addition of a single nucleotide (i.e. without being limited by theory, via the activity of TdT) lead to a further increase in CDRL3 length. This effect can be considered by introducing variants for the L3-Vλ sequences contributed by these four IGλV sequences (Table 38).

Table 38. Variants with an additional residue in CDRL3

Name	Locus	FRM1	CDR1	FRM2	CDR2	FRM3	CDR3 / L3-Vλ
1E+	IGVλ1-40+	QSVLTQPPSVSGAPGQRVTISC	TGSSSNIGAGYD---VH	WYQQLPGTAPKLLI	YGN----SNRPS	GVPDRFSGSKSG--TSASLAITGLQAEDEADYYC	QSYDSSLSG**S**
3L+	IGVλ3-19+	SSELTQDPAVSVALGQTVRITC	QGDSLRSYY------AS	WYQQKPGQAPVLVI	YGK----NNRPS	GIPDRFSGSSSG--NTASLTITGAQAEDEADYYC	NSRDSSGNH**H/Q**
3H+	IGVλ3-21+	SYVLTQPPSVSVAPGKTARITC	GGNNIGSKS------VH	WYQQKPGQAPVLVI	YYD----SDRPS	GIPERFSGSNSG--NTATLTISRVEAGDEADYYC	QVWDSSSDH**P**
6A+	IGVλ6-57+	NFMLTQPHSVSESPGKTVTISC	TRSSGSIASNY----VQ	WYQQRPGSSPTTVI	YED----NQRPS	GVPDRFSGSIDSSSNSASLTISGLKTEDEADYYC	QSYDSSN**H/Q**--

(+) sequences are derived from their parents by the addition of an amino acid at the end of the respective CDR3 (**bold underlined**). H/Q can be introduced in a single sequence by use of the degenerate codon CAW or similar.

Thus, the final set of chassis in the currently exemplified embodiment of the invention is 15: eleven contributed by the chassis in Table 14 and an additional four contributed by the chassis of Table 38. The corresponding L3-Vλ domains of the 15 chassis contribute from 7 to 10 amino acids to CDRL3. When considering the amino acids contributed by the IGλJ sequences, the total variation in the length of CDRL3 is 8 to 12 amino acids, approximating the distribution in Figure 4. Thus, in this exemplary embodiment of the invention, the minimalist Vλ library may be represented by the following: 15 Chassis × 5 IGλJ-derived segments = 75 sequences. Here, the 15 chassis are Vλ1-40, Vλ1-44, Vλ1-51, Vλ2-14, Vλ3-1*, Vλ3-19, Vλ3-21, Vλ4-69, Vλ6-57, Vλ5-45, Vλ7-43, Vλ1-40+, Vλ3-19+, Vλ3-21+, and Vλ6-57+. The 5 IGλJ-derived segments are YVFGGGTKLTVL (IGλJ1), VVFGGGTKLTVL (IGλJ2), WVFGGGTKLTVL (IGλJ3), AVFGGGTKLTVL (IGλJ), and -VFGGGTKLTVL (from any of the preceding sequences).

Example 8: Matching to "Reference" Antibodies

CDRH3 sequences of human antibodies of interest that are known in the art, (*e.g.,* antibodies that have been used in the clinic) have close counterparts in the designed library of the invention. A set of fifteen CDRH3 sequences from clinically relevant antibodies is presented in Table 39.

Table 39. CDRH3 Sequences of Reference Antibodies

Antibody Name	**Target**	**Origin**	**Status**	**CDHR3 sequence**	**SEQ ID NO:**
CAB1	TNF-α	Phage display - human library	FDA Approved	AKVSYLSTASSLDY	
CAB2	EGFR	Transgenic mouse	FDA Approved	VRDRVTGAFDI	
CAB3	IL-12/IL-23	Phage display - human library	Phase III	KTHGSHDN	
CAB4	Interleukin-1-β	Transgenic mouse	Phase III	ARDLRTGPFDY	
CAB5	RANKL	Transgenic mouse	Phase III	AKDPGTTVIMSWFDP	

CAB6	IL-12/IL-23	Transgenic mouse	Phase III	ARRRPGQGYFDF	
CAB7	TNF-α	Transgenic mouse	Phase III	ARDRGASAGGNYYYYGMDV	
CAB8	CTLA4	Transgenic mouse	Phase III	ARDPRGATLYYYYYGMDV	
CAB9	CD20	Transgenic mouse	Phase III	AKDIQYGNYYYGMDV	
CAB10	CD4	Transgenic mouse	Phase III	ARVINWFDP	
CAB11	CTLA4	Transgenic mouse	Phase III	ARTGWLGPFDY	
CAB12	IGF1-R	Transgenic mouse	Phase II	AKDLGWSDSYYYYYGMDV	
CAB13	EGFR	Transgenic mouse	Phase II	ARDGITMVRGVMKDYFDY	
CAB14	EGFR	Phage display - human library	Phase II	ARVSIFGVGTFDY	
CAB15	BLyS	Phage display - human library	Phase II	ARSRDLLLFPHHALSP	

Each of the above sequences was compared to each of the members of the library of Example 5, and the member, or members, with the same length and fewest number of amino acid mismatches was, or were, recorded. The results are summarized in Table 40, below. For most of the cases, matches with 80% identity or better were found in the exemplified CDRH3 library. To the extent that the specificity and binding affinity of each of these antibodies is influenced by their CDRH3 sequence, without being bound by theory, one or more of these library members could have measurable affinity to the relevant targets.

Table 40. Match of Reference Antibody CDRH3 to Designed Library

Antibody Name	**Number of Mismatches (*)**	**Length**	**% Identity of Best Match**
CAB1	5	14	64%
CAB2	2	11	82%
CAB3	4	8	50%
CAB4	2	11	82%
CAB5	3	15	80%
CAB6	3	12	75%
CAB7	2	20	90%
CAB8	0	19	100%
CAB9	3	15	80%
CAB10	1	9	89%

CAB11	1	11	91%
CAB12	2	18	89%
CAB13	2	18	89%
CAB14	1	13	92%
CAB15	7	16	56%

(*) For the best-matching sequence(s) in library

Given that a physical realization of a library with about 10^8 distinct members could, in practice, contain every single member, then such sequences with close percent identity to antibodies of interest would be present in the physical realization of the library. This example also highlights one of many distinctions of the libraries of the current invention over those of the art; namely, that the members of the libraries of the invention may be precisely enumerated. In contrast, CDRH3 libraries known in the art cannot be explicitly enumerated in the manner described herein. For example, many libraries known in the art (*e.g.,* Hoet *et al.,* Nat. Biotechnol., 2005, 23: 344; Griffiths *et al.,* EMBO J., 1994, 13: 3245; Griffiths *et al.,* EMBO J., 1993, 12: 725; Marks *et al.,* J. Mol. Biol., 1991, 222: 581, each incorporated by reference in its entirety) are derived by cloning of natural human CDRH3 sequences and their exact composition is not characterized, which precludes enumeration.

Synthetic libraries produced by other (*e.g.,* random or semi-random / biased) methods (Knappik, *et al.*, J Mol Biol, 2000, 296: 57, incorporated by reference in its entirety) tend to have very large numbers of unique members. Thus, while matches to a given input sequence (for example, at 80% or greater) may exist in a theoretical representation of such libraries, the probability of synthesizing and then producing a physical realization of the theoretical library that contains such a sequence and then selecting an antibody corresponding to such a match, in practice, may be remotely small. For example, a CDRH3 of length 19 in the Knappik library may have over 10^{19} distinct sequences. In a practical realization of such a library a tenth or so of the sequences may have length 19 and the largest total library may have in the order of 10^{10} to 10^{12} transformants; thus, the probability of a given pre-defined member being present, in practice, is effectively zero (less than one in ten million). Other libraries (*e.g.,* Enzelberger *et al.* WO2008053275 and Ladner US20060257937, each incorporated by reference in its entirety) suffer from at least one of the limitations described throughout this application.

Thus, for example, considering antibody CAB14, there are seven members of the designed library of Example 5 that differ at just one amino acid position from the sequence of the CDRH3 of CAB14 (given in Table 39). Since the total length of this CDRH3 sequence is 13, the percent of identical amino acids is 12/13 or about 92% for each of these 7 sequences of the library of the invention. It can be estimated that the probability of obtaining such a match (or better) in the library of Knappik *et al.* is about 1.4×10^{-9}; it would be lower still, about 5.5×10^{-10}, in a library with equal amino acid proportions (*i.e.,* completely random). Therefore, in a physical realization of the library with about 10^{10} transformants of which about a tenth may have length 13, there may be one or two instances of these best matches. However, with longer sequences such as CAB12, the probability of having members in the Knappik library with about 89% or better matching are under about 10^{-15}, so that the expected number of instances in a physical realization of the library is essentially zero. To the extent that sequences of interest resemble actual human CDRH3 sequences, there will be close matches in the library of Example 5, which was designed to mimic human sequences. Thus, one of the many relative advantages of the present library, versus those in the art, becomes more apparent as the length of the CDRH3 increases.

Example 9: Split Pool Synthesis of Oligonucleotides Encoding the DH, N2, and H3-JH Segments

This example outlines the procedures used to synthesize the oligonucleotides used to construct the exemplary libraries of the invention. Custom Primer Support ™ 200 dT40S resin (GE Healthcare) was used to synthesize the oligonucleotides, using a loading of about 39 μmol/g of resin. Columns (diameter = 30 μm) and frits were purchased from Biosearch Technologies, Inc. A column bed volume of 30 μL was used in the synthesis, with 120 nmol of resin loaded in each column. A mixture of dichloromethane (DCM) and methanol (MeOH), at a ratio of 400/122 (v/v) was used to load the resin. Oligonucleotides were synthesized using a Dr. Oligo ® 192 oligonucleotide synthesizer and standard phosphorothioate chemistry.

The split pool procedure for the synthesis of the [DH]-[N2]-[H3-JH] oligonucleotides was performed as follows: First, oligonucleotide leader sequences, containing a randomly chosen 10 nucleotide sequence (ATGCACAGTT; SEQ ID NO:___), a BsrDI recognition site (GCAATG), and a two base "overlap sequence" (TG, AC, AG, CT, or GA) were synthesized. The purpose of each of these segments is explained below. After synthesis of this 18 nucleotide sequence, the DH segments were synthesized; approximately 1 g of resin (with the 18 nucleotide segment still conjugated) was suspended in 20 mL of DCM/MeOH. About 60 μL of the resulting slurry (120 nmol) was distributed inside each of 278 oligonucleotide synthesis columns. These 278 columns were used to synthesize the 278 DH segments of Table 18, 3' to the 18 nucleotide segment described above. After synthesis, the 278 DH segments were pooled as follows: the resin and frits were pushed out of the columns and collected inside a 20 mL syringe barrel (without plunger). Each column was then washed with 0.5 mL MeOH, to remove any residual resin that was adsorbed to the walls of the column. The resin in the syringe barrel was washed three times with MeOH, using a low porosity glass filter to retain the resin. The resin was then dried and weighed.

The pooled resin (about 1.36 g) containing the 278 DH segments was subsequently suspended in about 17 mL of DCM/MeOH, and about 60 μL of the resulting slurry was distributed inside each of two sets of 141 columns. The 141 N2 segments enumerated in Tables 24 and 25 were then synthesized, in duplicate (282 total columns), 3' to the 278 DH segments synthesized in the first step. The resin from the 282 columns was then pooled, washed, and dried, as described above.

The pooled resin obtained from the N2 synthesis (about 1.35 g) was suspended in about 17 mL of DCM/MeOH, and about 60 μL of the resulting slurry was distributed inside each of 280 columns, representing 28 H3-JH segments synthesized ten times each. A portion (described more fully below) of each of the 28 IGHJ segments, including H3-JH of Table 20 were then synthesized, 3' to the N2 segments, in ten of the columns. Final oligonucleotides were cleaved and deprotected by exposure to gaseous ammonia (85°C, 2 h, 60 psi).

Split pool synthesis was used to synthesize the exemplary CDRH3 library. However, it is appreciated that recent advances in oligonucleotide synthesis, which enable the synthesis of longer oligonucleotides at higher fidelity and the production of the oligonucleotides of the library by synthetic procedures that involve splitting, but not pooling, may be used in alternative embodiments of the invention. The split pool synthesis described herein is, therefore, one possible means of obtaining the oligonucleotides of the library, but is not limiting. One other possible means of synthesizing the oligonucleotides described in this application is the use of trinucleotides. This may be expected to increase the fidelity of the synthesis, since frame shift mutants would be reduced or eliminated.

Example 10: Construction of the CDRH3 and Heavy Chain Libraries

This example outlines the procedures used to create exemplary CDRH3 and heavy chain libraries of the invention. A two step process was used to create the CDRH3 library. The first step involved the assembly of a set of vectors encoding the tail and N1 segments, and the second step involved utilizing the split pool nucleic acid synthesis procedures outlined in Example 9 to create oligonucleotides encoding the DH, N2, and H3-JH segments. The chemically synthesized oligonucleotides were then ligated into the vectors, to yield CDRH3 residues 95-102, based on the numbering system described herein. This CDRH3 library was subsequently amplified by PCR and recombined into a plurality of vectors containing the heavy chain chassis variants described in Examples 1 and 2. CDRH1 and CDRH2 variants were produced by QuikChange ® Mutagenesis (Stratagene ™), using the oligonucleotides encoding the ten heavy chain chassis of Example 1 as a template. In addition to the heavy chain chassis, the plurality of vectors contained the heavy chain constant regions (*i.e.,* CH1, CH2, and CH3) from IgG1, so that a full-length heavy chain was formed upon recombination of the CDRH3 with the vector containing the heavy chain chassis and constant regions. In this exemplary embodiment, the recombination to produce the full-length heavy chains and the expression of the full-length heavy chains were both performed in *S. cerevisiae.*

To generate full-length, heterodimeric IgGs, comprising a heavy chain and a light chain, a light chain protein was also expressed in the yeast cell. The light chain library used in this embodiment was the kappa light chain library, wherein the VKCDR3s were synthesized using degenerate oligonucleotides (see Example 6.2). Due to the shorter length of the oligonucleotides encoding the light chain library (in comparison to those encoding the heavy chain library), the light chain CDR3 oligonucleotides could be synthesized *de novo*, using standard procedures for oligonucleotide synthesis, without the need for assembly from sub-components (as in the heavy chain CDR3 synthesis). One or more light chains can be expressed in each yeast cell which expresses a particular heavy chain clone from a library of the invention. One or more light chains have been successfully expressed from both episomal (*e.g.,* plasmid) vectors and from integrated sites in the yeast genome.

Below are provided further details on the assembly of the individual components for the synthesis of a CDRH3 library of the invention, and the subsequent combination of the exemplary CDRH3 library with the vectors containing the chassis and constant regions. In this particular exemplary embodiment of the invention, the steps involved in the process may be generally characterized as (i) synthesis of 424 vectors encoding the tail and N1 regions; (ii) ligation of oligonucleotides encoding the [DH]-[N2]-[H3-JH] segments into these 424 vectors; (iii) PCR amplification of the CDRH3 sequences from the vectors produced in these ligations; and (iv) homologous recombination of these PCR-amplified CDRH3 domains into the yeast expression vectors containing the chassis and constant regions.

Example 10.1: Synthesis of Vectors Encoding the Tail and N1 Regions

This example demonstrates the synthesis of 424 vectors encoding the tail and N1 regions of CDRH3. In this exemplary embodiment of the invention, the tail was restricted to G, D, E, or nothing, and the N1 region was restricted to one of the 59 sequences shown in Table 24. As described throughout the specification, many other embodiments are possible.

In the first step of the process, a single "base vector" (pJM204, a pUC-derived cloning vector) was constructed, which contained (i) a nucleic acid sequence encoding two amino acids that are common to the C-terminal portion of all 28 IGHJ segments (SS), and (ii) a nucleic acid sequence encoding a portion of the CH1 constant region from IgG1. Thus, the base vector contains an insert encoding a sequence that can be depicted as:

[SS]-[CH1~],

wherein SS is a common portion of the C-terminus of the 28 IGHJ segments and CH1~ is a portion of the CH1 constant region from IgG1, namely: ASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLG (SEQ ID NO:___).

Next, 424 different oligonucleotides were cloned into the base vector, upstream (*i.e.,* 5') from the region encoding the [SS]-[CH1~]. These 424 oligonucleotides were synthesized by standard methods and each encoded a C-terminal portion of one of the 17 heavy chain chassis enumerated in Table 5, plus one of four exemplary tail segments (G/D/E/-), and one of 59 exemplary N1 segments (Table 24). These 424 oligonucleotides, therefore, encode a plurality of sequences that may be represented by:

[~FRM3]-[G/D/E/-]-[N1],

wherein ~FRM3 represents a C-terminal portion of a FRM3 region from one of the 17 heavy chain chassis of Table 5, G/D/E/- represents G, D, E, or nothing, and N1 represents one of the 59 N1 sequences enumerated in Table 24. As described throughout the specification, the invention is not limited to the chassis exemplified in Table 5, their CDRH1 and CDRH2 variants (Table 8), the four exemplary tail options used in this example, or the 59 N1 segments presented in Table 24.

The oligonucleotide sequences represented by the sequences above were synthesized in two groups: one group containing a ~FRM3 region identical to the corresponding region on 16 of the 17 the heavy chain chassis enumerated in Table 5, and another group containing a ~FRM3 region that is identical to the corresponding region on VH3-15. In the former group, an oligonucleotide encoding DTAVYYCA<u>R</u> (SEQ ID NO:___) was used for ~FRM3. During subsequent PCR amplification, the V residue of VH5-51 was altered to an M, to correspond to the VH5-51 germline sequence. In the latter group (that with a sequence common to VH3-15), a larger oligonucleotide, encoding the sequence AISGSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCA<u>K</u> (SEQ ID NO:___) was used for ~FRM3. Each of the two oligonucleotides encoding the ~FRM3 regions were paired with oligonucleotides encoding one of the four tail regions (G/D/E/-) and one of the 59 N1 segments, yielding a total of 236 possible combinations for each ~FRM3 (*i.e.,* 1 × 4 × 59), or a total of 472 possible combinations when both ~FRM3 sequences are considered. However, 48 of these combinations are redundant and only a single representation of these sequences was used in the currently exemplified CDRH3 library, yielding 424 unique oligonucleotides encoding [~FRM3]-[G/D/E/-]-[N1] sequences.

After the oligonucleotides encoding the [~FRM3]-[G/D/E/-]-[N1] and [SS]-[CH1~] segments were cloned into the vector, as described above, additional sequences were added to the vector to facilitate the subsequent insertion of the oligonucleotides encoding the [DH]-[N2]-[H3-JH] fragments synthesized during the split pool synthesis. These additional sequences comprise a polynucleotide encoding a selectable marker protein, flanked on each side by a recognition site for a type II restriction enzyme, for example:

[Type II RS 1]-[selectable marker protein]-[Type II RS 2].

In this exemplary embodiment, the selectable marker protein is ccdB and the type II restriction enzyme recognition sites are specific for BsrDI and BbsI. In certain strains of *E. coli*, the ccdB protein is toxic, thereby preventing the growth of these bacteria when the gene is present.

An example of the 5’ end of one of the 212 vectors with a ~FRM3 region based on the VH3-23 chassis, D tail residue and an N1 segment of length zero is presented below:

```
                                                                              VH3-23
                                                                   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                                    A  I  S   G  S  G   G  S  T  Y ·
 961                                                               GCTATTAG TGGTAGTGGT GGTAGCACAT
                                                                   CGATAATC ACCATCACCA CCATCGTGTA

                                                  VH3-23
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          ·  Y  A  D   S  V  K   G  R  F   T  I  S  R   D  N  S   K  N  T  L   Y  L   Q  M  N  S
1041      ACTACGCAGA CTCCGTGAAG GGCCGGTTCA CCATCTCCAG AGACAATTCC AAGAACACGC TGTATCTGCA AATGAACAGC
          TGATGCGTCT GAGGCACTTC CCGGCCAAGT GGTAGAGGTC TCTGTTAAGG TTCTTGTGCG ACATAGACGT TTACTTGTCG

                          VH3-23                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~                 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                        BsrDI
                                                        ~~~~~~
           L  R  A  E   D  T  A   V  Y  Y   C  A  K
1121      CTGAGAGCCG AGGACACGGC GGTGTACTAC TGCGCCAAGG ACCATTGCGC TTAGCCTAGG TTATATTCCC CAGAACATCA
          GACTCTCGGC TCCTGTGCCG CCACATGATG ACGCGGTTCC TGGTAACGCG AATCGGATCC AATATAAGGG GTCTTGTAGT
```

An example of one of the 212 vectors with a ~FRM3 region based on one of the other 16 chassis, with a D residue as the tail and an N1 segment of length zero is presented below:

```
                                                                          Framework 3
                                                                  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                                   D  T  A   V  Y  Y  C   A  R
 961                                                              GACACGGCG GTGTACTACT GCGCCAGAGA
                                                                  CTGTGCCGC CACATGATGA CGCGGTCTCT
                                            ccdB
           ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
           BsrDI
           ~~~~~~
1041      CCATTGCGCT TAGCCTAGGT TATATTCCCC AGAACATCAG GTTAATGGCG TTTTTGATGT CATTTTCGCG GTGGCTGAGA
          GGTAACGCGA ATCGGATCCA ATATAAGGGG TCTTGTAGTC CAATTACCGC AAAAACTACA GTAAAAGCGC CACCGACTCT
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1121      TCAGCCACTT CTTCCCCGAT AACGGAAACC GGCACACTGG CCATATCGGT GGTCATCATG CGCCAGCTTT CATCCCCGAT
          AGTCGGTGAA GAAGGGGCTA TTGCCTTTGG CCGTGTGACC GGTATAGCCA CCAGTAGTAC GCGGTCGAAA GTAGGGGCTA
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1201      ATGCACCACC GGGTAAAGTT CACGGGAGAC TTTATCTGAC AGCAGACGTG CACTGGCCAG GGGGATCACC ATCCGTCGCC
          TACGTGGTGG CCCATTTCAA GTGCCCTCTG AAATAGACTG TCGTCTGCAC GTGACCGGTC CCCCTAGTGG TAGGCAGCGG
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1281      CGGGCGTGTC AATAATATCA CTCTGTACAT CCACAAACAG ACGATAACGG CTCTCTCTTT TATAGGTGTA AACCTTAAAC
          GCCCGCACAG TTATTATAGT GAGACATGTA GGTGTTTGTC TGCTATTGCC GAGAGAGAAA ATATCCACAT TTGGAATTTG
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1361      TGCATTTCAC CAGCCCCTGT TCTCGTCAGC AAAAGAGCCG TTCATTTCAA TAAACCGGGC GACCTCAGCC ATCCCTTCCT
          ACGTAAAGTG GTCGGGGACA AGAGCAGTCG TTTTCTCGGC AAGTAAAGTT ATTTGGCCCG CTGGAGTCGG TAGGGAAGGA
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1441      GATTTTCCGC TTTCCAGCGT TCGGCACGCA GACGACGGGC TTCATTCTGC ATGGTTGTGC TTACCAGACC GGAGATATTG
          CTAAAAGGCG AAAGGTCGCA AGCCGTGCGT CTGCTGCCCG AAGTAAGACG TACCAACACG AATGGTCTGG CCTCTATAAC
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1521      ACATCATATA TGCCTTGAGC AACTGATAGC TGTCGCTGTC AACTGTCACT GTAATACGCT GCTTCATAGC ATACCTCTTT
          TGTAGTATAT ACGGAACTCG TTGACTATCG ACAGCGACAG TTGACAGTGA CATTATGCGA CGAAGTATCG TATGGAGAAA
                                            ccdB
          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
1601      TTGACATACT TCGGGTATAC ATATCAGTAT ATATTCTTAT ACCGCAAAAA TCAGCGCGCA AATATGCATA CTGTTATCTG
```

```
        AACTGTATGA AGCCCATATG TATAGTCATA TATAAGAATA TGGCGTTTTT AGTCGCGCGT TTATACGTAT GACAATAGAC
               ccdB                                        CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                      BbsI
                                     ~~~~~~~
                                                   A   S  T  K   G  P  S   V  F  P  L   A  P  S ·
1681    GCTTTTAGTA AGCCGCCTAG GTCATCAGAA GACAACTCAG CTAGCACCAA GGGCCCATCG GTCTTTCCCC TGGCACCCTC
        CGAAAATCAT TCGGCGGATC CAGTAGTCTT CTGTTGAGTC GATCGTGGTT CCCGGGTAGC CAGAAAGGGG ACCGTGGGAG
                                              CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        · S  K  S   T  S  G  G   T  A  A   L  G  C   L  V  K  D   Y  F  P   E  P  V   T  V  S  W ·
1761    CTCCAAGAGC ACCTCTGGGG GCACAGCGGC CCTGGGCTGC CTGGTCAAGG ACTACTTCCC CGAACCGGTG ACGGTGTCGT
        GAGGTTCTCG TGGAGACCCC CGTGTCGCCG GGACCCGACG GACCAGTTCC TGATGAAGGG GCTTGGCCAC TGCCACAGCA
                                       CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ·  N  S  G   A  L  T   S  G  V  H   T  F  P   A  V  L   Q  S  S  G   L
1841    GGAACTCAGG CGCCCTGACC AGCGGCGTGC ACACCTTCCC GGCTGTCCTA CAGTCCTCAG GACTC
        CCTTGAGTCC GCGGGACTGG TCGCCGCACG TGTGGAAGGG CCGACAGGAT GTCAGGAGTC CTGAG
```

All 424 vectors were sequence verified. A schematic diagram of the content of the 424 vectors, before and after cloning of the [DH]-[N2]-[H3-JH] fragment is presented in Figure 5. Below is an exemplary sequence from one of the 424 vectors containing a FRM3 region from VH3-23.

```
                                                              primer EMK135
                                                          ~~~~~~~~~~~~~~~~~~~~~~~~
                                                          VH3-23
                                 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                  A  I  S   G  S  G   G  S  T   Y  Y  A  D   S  V  K   G  R  F
561                              GCTATTA GTGGTAGTGG TGGTAGCACA TACTACGCAG ACTCCGTGAA GGGCCGGTTC
                                 CGATAAT CACCATCACC ACCATCGTGT ATGATGCGTC TGAGGCACTT CCCGGCCAAG
                                             VH3-23
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         T  I  S  R   D  N  S   K  N  T   L  Y  L  Q   M  N  S   L  R  A   E  D  T  A   V  Y  Y ·
641     ACCATCTCCA GAGACAATTC CAAGAACACG CTGTATCTGC AAATGAACAG CCTGAGAGCC GAGGACACGG CGGTGTACTA
        TGGTAGAGGT CTCTGTTAAG GTTCTTGTGC GACATAGACG TTTACTTGTC GGACTCTCGG CTCCTGTGCC GCCACATGAT
          VH3-23                              D                                    J1
        ~~~~~~~~~~~             ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~          ~~~~~~~~~~~~~~~~~~~~
                                                                                   JH6
                                                                           ~~~~~~~~~~~~~~~~~~~~
                       N1_9                                         N2
                   ~~~~~~~~~~~~~                                 ~~~~~~~~~~
        · C  A  K   D  A  G  G   Y  Y  Y   G  S  G   S  Y  Y  N   A  A  A   Y  Y  Y   Y  Y  G  M ·
721     CTGCGCCAAG GACGCCGGAG GATATTATTA TGGGTCAGGA AGCTATTACA ACGCTGCGGC TTACTACTAC TATTATGGCA
        GACGCGGTTC CTGCGGCCTC CTATAATAAT ACCCAGTCCT TCGATAATGT TGCGACGCCG AATGATGATG ATAATACCGT
                           JH6
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
           J1                                                         CH1
        ~~~~~~~~                                     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                      NheI
                                                     ~~~~~~
        ·  D  V  W   G  Q  G   T  T  V  T   V  S  S   A  S  T   K  G  P  S   V  F  P   L  A  P
801     TGGACGTGTG GGGACAAGGT ACAACAGTCA CCGTCTCCTC AGCTAGCACC AAGGGCCCAT CGGTCTTTCC CCTGGCACCC
        ACCTGCACAC CCCTGTTCCA TGTTGTCAGT GGCAGAGGAG TCGATCGTGG TTCCCGGGTA GCCAGAAAGG GGACCGTGGG
                                              CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         S  S  K  S   T  S  G   G  T  A   A  L  G  C   L  V  K   D  Y  F   P  E  P  V   T  V  S ·
881     TCCTCCAAGA GCACCTCTGG GGGCACAGCG GCCCTGGGCT GCCTGGTCAA GGACTACTTC CCCGAACCGG TGACGGTGTC
        AGGAGGTTCT CGTGGAGACC CCCGTGTCGC CGGGACCCGA CGGACCAGTT CCTGATGAAG GGGCTTGGCC ACTGCCACAG
                                               EK137 CH1 Primer
                                              ~~~~~~~~~~~~~~~~~~~~
                                              CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        · W  N  S   G  A  L  T   S  G  V   H  T  F   P  A  V  L   Q  S  S   G  L  Y   S  L  S  S ·
961     GTGGAACTCA GGCGCCCTGA CCAGCGGCGT GCACACCTTC CCGGCTGTCC TACAGTCCTC AGGACTCTAC TCCCTCAGCA
        CACCTTGAGT CCGCGGGACT GGTCGCCGCA CGTGTGGAAG GGCCGACAGG ATGTCAGGAG TCCTGAGATG AGGGAGTCGT
                      CH1
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ·  V  V  T   V  P  S   S  S  L  G
1041    GCGTGGTGAC CGTGCCCTCC AGCAGCTTGG GC
        CGCACCACTG GCACGGGAGG TCGTCGAACC CG
```

Example 10.2: Cloning of the Oligonucleotides Encoding the DH, N2, H3-JH Segments into the Vectors Containing the Tail and N1 Segments

This example describes the cloning of the oligonucleotides encoding the [D]-[N2]-[H3-JH] segments (made via split pool synthesis; Example 9) into the 424 vectors produced in Example 10.1. To summarize, the [DH]-[N2]-[H3-JH] oligonucleotides produced via split pool synthesis were amplified by PCR, to produce double-stranded oligonucleotides, to introduce restriction sites that would create overhangs complementary to those on the vectors (*i.e.,* BsrDI and BbsI), and to complete the 3' portion of the IGHJ segments that was not synthesized in the split pool synthesis. The amplified oligonucleotides were then digested with the restriction enzymes BsrDI (cleaves adjacent to the DH segment) and BbsI (cleaves near the end of the JH segment). The cleaved oligonucleotides were then purified and ligated into the 424 vectors which had previously been digested with BsrDI and BbsI. After ligation, the reactions were purified, ethanol precipitated, and resolubilized.

This process for one of the [DH]-[N2]-[H3-JH] oligonucleotides synthesized in the split pool synthesis is illustrated below. The following oligonucleotide (SEQ ID NO:___) is one of the oligonucleotides synthesized during the split pool synthesis:

1 ATGCACAGTTGCAATG**TG**TATTACTATGGATCTGGTTCTTACTATAAT**GT** 50
51 **GGGCGGA**TATTATTACTACTATGGTATGGACGTATGGGGGCAAGGGACC 99

The first 10 nucleotides (ATGCACAGTT; SEQ ID NO:___) represent a portion of a random sequence that is increased to 20 base pairs in the PCR amplification step, below. This portion of the sequence increases the efficiency of BsrDI digestion and facilitates the downstream purification of the oligonucleotides.

Nucleotides 11-16 (underlined) represent the BsrDI recognition site. The two base overlap sequence that follows this site (in this example TG; bold) was synthesized to be complementary to the two base overhang created by digesting certain of the 424 vectors with BsrDI (*i.e.,* depending on the composition of the tail / N1 region of the particular vector). Other oligonucleotides contain different two-base overhangs, as described below.

The two base overlap is followed by the DH gene segment (nucleotides 19-48), in this example, by a 30 bp sequence (TATTACTATGGATCTGGTTCTTACTATAAT, SEQ ID NO:___) which encodes the ten residue DH segment YYYGSGSYYN (*i.e.,* IGHD3-10_2 of Table 17; SEQ ID NO:___).

The region of the oligonucleotide encoding the DH segment is followed, in this example, by a nine base region (GTGGGCGGA; bold; nucleotides 49-57), encoding the N2 segment (in this case VGG; Table 24).

The remainder of this exemplary oligonucleotide represents the portion of the JH segment that is synthesized during the split pool synthesis (TATTATTACTACTATGGTATGGACGTATGGGGGCAAGGGACC; SEQ ID NO: ___; nucleotides 58-99; underlined), encoding the sequence YYYYYGMDVWGQGT (Table 20; SEQ ID NO:___). The balance of the IGHJ segment is added during the subsequent PCR amplification described below.

After the split pool-synthesized oligonucleotides were cleaved from the resin and deprotected, they served as a template for a PCR reaction which added an additional randomly chosen 10 nucleotides (*e.g.,* GACGAGCTTC; SEQ ID NO:___) to the 5' end and the rest of the IGHJ segment plus the BbsI restriction site to the 3' end. These additions facilitate the cloning of the [DH]-[N2]-[JH] oligonucleotides into the 424 vectors. As described above (Example 9), the last round of the split pool synthesis involves 280 columns: 10 columns for each of the oligonucleotides encoding one of 28 H3-JH segments. The oligonucleotide products obtained from these 280 columns are pooled according to the identity of their H3-JH segments, for a total of 28 pools. Each of these 28 pools is then amplified in five separate PCR reactions, using five forward primers that each encode a different two base overlap (preceding the DH segment; see above) and one reverse primer that has a sequence corresponding to the familial origin of the H3-JH segment being amplified. The sequences of these 11 primers are provided below:

Forward primers

AC	GACGAGCTTCAATGCACAGTTGCAATGAC	(SEQ ID NO:___)
AG	GACGAGCTTCAATGCACAGTTGCAATGAG	(SEQ ID NO:___)
CT	GACGAGCTTCAATGCACAGTTGCAATGCT	(SEQ ID NO:___)
GA	GACGAGCTTCAATGCACAGTTGCAATGGA	(SEQ ID NO:___)
TG	GACGAGCTTCAATGCACAGTTGCAATGTG	(SEQ ID NO:___)

Reverse Primers

JH1	TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACGGTGACCAAGGTGCCCTGGCCCCA (SEQ ID NO:___)
JH2	TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACAGTGACCAAGGTGCCACGGCCCCA (SEQ ID NO:___)
JH3	TGCATCAGTGCGACTAACGGAAGACTCTGAAGAGACGGTGACCATTGTCCCTTGGCCCCA (SEQ ID NO:___)
JH4	TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACGGTGACCAAGGTTCCTTGGCCCCA

(SEQ ID NO:___)
JH5 TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACGGTGACCAAGGTTCCCTGGCCCCA
(SEQ ID NO:___)
JH6 TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACGGTGACCGTGGTCCCTTGCCCCCA
(SEQ ID NO:___)

Amplifications were performed using Taq polymerase, under standard conditions. The oligonucleotides were amplified for eight cycles, to maintain the representation of sequences of different lengths. Melting of the strands was performed at 95°C for 30 seconds, with annealing at 58°C and a 15 second extension time at 72°C.

Using the exemplary split-pool derived oligonucleotide enumerated above as an example, the PCR amplification was performed using the TG primer and the JH6 primer, where the annealing portion of the primers has been underlined:

TG GACGAGCTTCAATGCACAGTTGCAATGTG (SEQ ID NO:___)

JH6 TGCATCAGTGCGACTAACGGAAGACTCTGAGGAGACGGTGACCGTGGTCCCTTGCCCCCA
(SEQ ID NO:___)

The portion of the TG primer that is 5' to the annealing portion includes the random 10 base pairs described above. The portion of the JH6 primer that is 5' to the annealing portion includes the balance of the JH6 segment and the BbsI restriction site. The following PCR product (SEQ ID NO:___) is formed in the reaction (added sequences underlined):

GACGAGCTTCATGCACAGTTGCAATGTGTATTACTATGGATCTGGTTCTTACTATAATGTGGGCGGATATTAT
TACTACTATGGTATGGACGTATGGGGGCAAGGGACCACGGTCACCGTCTCCTCAGAGTCTTCCGTTAGTCGCA
CTGATGCAG

The PCR products from each reaction were then combined into five pools, based on the forward primer that was used in the reaction, creating sets of sequences yielding the same two-base overhang after BsrDI digestion. The five pools of PCR products were then digested with BsRDI and BbsI (100 µg of PCR product; 1 mL reaction volume; 200 U BbsI; 100 U BsrDI; 2h; 37°C; NEB Buffer 2). The digested oligonucleotides were extracted twice with phenol/chloroform, ethanol precipitated, air dried briefly and resolubilized in 300 µL of TE buffer by sitting overnight at 4°C.

Each of the 424 vectors described in the preceding sections was then digested with BsrDI and BbsI, each vector yielding a two base overhang that was complimentary to one of those contained in one of the five pools of PCR products. Thus, one of the five pools of restriction digested PCR products are ligated into each of the 424 vectors, depending on their compatible ends, for a total of 424 ligations.

Example 10.3: PCR Amplification of the CDRH3 from the 424 Vectors

This example describes the PCR amplification of the CDRH3 regions from the 424 vectors described above. As set forth above, the 424 vectors represent two sets: one for the VH3-23 family, with FRM3 ending in CAK (212 vectors) and one for the other 16 chassis, with FRM3 ending in CAR (212 vectors). The CDRH3s in the VH3-23-based vectors were amplified using a reverse primer (EK137; see Table 41) recognizing a portion of the CH1 region of the plasmid and the VH3-23-specific primer EK135 (see Table 41). Amplification of the CDRH3s from the 212 vectors with FRM3 ending in CAR was performed using the same reverse primer (EK137) and each of five FRM3-specific primers shown in Table 41 (EK139, EK140, EK141, EK143, and EK144). Therefore, 212 VH3-23 amplifications and 212 × 5 FRM3 PCR reactions were performed, for a total of 1,272 reactions. An additional PCR reaction amplified the CDRH3 from the 212 VH3-23-based vectors, using the EK 133 forward primer, to allow the amplicons to be cloned into the other 5 VH3 family member chassis while making the last three amino acids of these chassis CAK instead of the original CAR (VH3-23*). The primers used in each reaction are shown in Table 41.

Table 41. Primers Used for Amplification of CDRH3 Sequences

Primer No.	Compatible Chassis	Primer Sequence	SEQ ID NO
EK135	VH3-23	CACATACTACGCAGACTCCGTG	
EK133	VH3-48; VH3-7; VH3-15; VH3-30; VH3-33; VH3-23*	CAAATGAACAGCCTGAGAGCCGAGGACACGGCGGTGTACTACTG	
EK139	VH4-B; VH4-31; VH4-34; VH4-39; VH4-59;	AAGCTGAGTTCTGTGACCGCCGCAGACACGGCGGTGTACTACTG	

	VH4-61		
EK140	VH1-46; VH1-69	GAGCTGAGCAGCCTGAGATCTGAGGACACGGCGGTGTACTACTG	
EK141	VH1-2	GAGCTGAGCAGGCTGAGATCTGACGACACGGCGGTGTACTACTG	
EK143	VH5-51	CAGTGGAGCAGCCTGAAGGCCTCGGACACGGCGATGTACTACTG	
EK144	VH1-18	GAGCTGAGGAGCCTGAGATCTGACGACACGGCGGTGTACTACTG	
EK137	CH1 Rev. Primer	GTAGGACAGCCGGGAAGG	

Example 10.4: Homologous Recombination of PCR-Amplified CDRH3 Regions Into Heavy Chain Chassis

After amplification, reaction products were pooled according to the respective VH chassis that they would ultimately be cloned into. Table 42 enumerates these pools, with the PCR primers used to obtain the CDRH3 sequences in each pool provided in the last two columns.

Table 42. PCR Primers Used to Amplify CDRH3 Regions from 424 Vectors

Pool # (Arbitrary)	HC Chassis Target	5' Primer	3' Primer
1	1-46	EK140	EK137
	1-69	EK140	EK137
2	1-2	EK141	EK137
3	1-18	EK144	EK137
4	4-B	EK139	EK137
	4-31	EK139	EK137
	4-34^2	EK139	EK137
	4-39	EK139	EK137
	4-59	EK139	EK137
	4-61	EK139	EK137
5	5-51	EK143	EK137
6	3-15^1	EK133	EK137
	3-7	EK133	EK137
	3-33	EK133	EK137
	3-33	EK133	EK137
	3-48	EK133	EK137
7	3-23	EMK135	EK137
8	3-23*	EK133	EK137

[*] Allowed the amplicons to be cloned into the other 5 VH3 family member chassis (*i.e.*, other than VH3-23), while making the last three amino acids of these chassis CAK instead of the original CAR.

[1] As described in Table 5, the original KT sequence in VH3-15 was mutated to RA, and the original TT to AR.

[2] As described in Table 5, the potential site for N-linked glycosylation was removed from CDRH2 of this chassis.

After pooling of the amplified CDRH3 regions, according to the process outlined above, the heavy chain chassis expression vectors were pooled according to their origin and cut, to create a "gap" for homologous recombination with the amplified CDRH3s. Figure 6 shows a schematic structure of a heavy chain vector, prior to recombination with a CDRH3. In this exemplary embodiment of the invention, there were a total of 152 vectors encoding heavy chain chassis and IgG1 constant regions, but no CDRH3. These 152 vectors represent 17 individual variable heavy chain gene families (Table 5; Examples 1 and 2). Fifteen of the families were represented by the heavy chain chassis sequences described in Table 5 and the CDRH1/H2 variants described in Table 8 (*i.e.*, 150 vectors). VH 3-30 differs from VH3-33 by a single amino acid; thus VH3-30 was included in the VH3-33 pool of variants. The 4-34 VH family member was kept separate from all others and, in this exemplary embodiment, no variants of it were included in the library. Thus, a total of 16 pools, representing 17 heavy chain chassis, were generated from the 152 vectors.

The vector pools were digested with the restriction enzyme SfiI, which cuts at two sites in the vector that are located between the end of the FRM3 of the variable domain and the start of the CH1.

```
                                              VH3-48
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
           S  V  K   G  R  F  T   I  S  R   D  N  A   K  N  S  L   Y  L  Q   M  N  S   L  R  A  E ·
2801     CTCTGTGAAG GGCCGATTCA CCATCTCCAG AGACAATGCC AAGAACTCAC TGTATCTGCA AATGAACAGC CTGAGAGCTG
         GAGACACTTC CCGGCTAAGT GGTAGAGGTC TCTGTTACGG TTCTTGAGTG ACATAGACGT TTACTTGTCG GACTCTCGAC
               Constant DTAVYYCAR
           ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         VH3-48                                                                            VTVSS common to all J
         ~~                                                                                ~~~~~
                                            SfiI                                 SfiI
                                        ~~~~~~~~~~~~~~~                      ~~~~~~~~~~~~~~
         ·  D  T  A   V  Y  Y   C  A  R
                                                                                            V  T ·
2881     AGGACACGGC GGTGTACTAC TGCGCCAGAG GCCAATAGGG CCAACTATAA CAGGGGTACC CCGGCCAATA AGGCCGTCAC
         TCCTGTGCCG CCACATGATG ACGCGGTCTC CGGTTATCCC GGTTGATATT GTCCCCATGG GGCCGGTTAT TCCGGCAGTG
         VTVSS common to all J
         ~~~~~~~~~~~
                                                  hIgG1m17,1
                    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                     NheI
                    ~~~~~~
         · V  S  S   A  S  T  K   G  P  S   V  F  P   L  A  P  S   S  K  S   T  S  G   G  T  A
2961     CGTCTCCTCA GCTAGCACCA AGGGCCCATC GGTCTTCCCC CTGGCACCCT CCTCCAAGAG CACCTCTGGG GGCACAGCGG
         GCAGAGGAGT CGATCGTGGT TCCCGGGTAG CCAGAAGGGG GACCGTGGGA GGAGGTTCTC GTGGAGACCC CCGTGTCGCC
```

The gapped vector pools were then mixed with the appropriate (*i.e.*, compatible) pool of CDRH3 amplicons, generated as described above, at a 50:1 insert to vector ratio. The mixture was then transformed into electrocompetent yeast (*S. cerevisiae*), which already contained plasmids or integrated genes comprising a VK light chain library (described below). The degree of library diversity was determined by plating a dilution of the electroporated cells on a selectable agar plate. In this exemplified embodiment of the invention, the agar plate lacked tryptophan and the yeast lacked the ability to endogenously synthesize tryptophan. This deficiency was remedied by the inclusion of the TRP marker on the heavy chain chassis plasmid, so that any yeast receiving the plasmid and recombining it with a CDRH3 insert would grow. The electroporated cells were then outgrown approximately 100-fold, in liquid media lacking tryptophan. Aliquots of the library were frozen in 50% glycerol and stored at -80°C. Each transformant obtained at this stage represents a clone that can express a full IgG molecule. A schematic diagram of a CDRH3 integrated into a heavy chain vector and the accompanying sequence are provided in Figure 5.

A heavy chain library pool was then produced, based on the approximate representation of the heavy chain family members as depicted in Table 43.

Table 43. Occurrence of Heavy Chain Chassis in Data Sets Used to Design Library, Expected (Designed) Library, and Actual (Observed) Library

Chassis	**Relative Occurrence in Data Sets (1)**	**Expected (2)**	**Observed (3)**
VH1-2	5.1	6.0	6.4
VH1-18	3.4	3.7	3.8
VH1-46	3.4	5.2	4.7
VH1-69	8.0	8.0	10.7
VH3-7	3.6	6.1	4.5
VH3-15	1.9	6.9	3.6
VH3-23	11.0	13.2	17.1
VH3-33/30	13.1	12.5	6.6
VH3-48	2.9	6.3	7.5
VH4-31	3.4	2.5	4.3
VH4-34	17.2	7.0	4.7
VH4-39	8.7	3.9	3.0
VH4-59	7.0	7.8	9.2
VH4-61	3.2	1.9	2.4
VH4-B	1.0	1.4	0.8
VH5-51	7.2	7.7	10.5

(1) As detailed in Example 1, these 17 sequences account for about 76% of the entire sample of human VH sequences used to represent the human repertoire.
(2) Based on pooling of sub-libraries of each chassis type.

(3) Usage in 531 sequences from library; cf. Figure 20.

Example 10.5: K94R Mutation in VH3-23 and R94K Mutation in VH3-33, VH3-30, VH3-7, and VH3-48

This example describes the mutation of position 94 in VH3-23, VH3-33, VH3-30, VH3-7, and VH3-48. In VH3-23, the amino acid at this position was mutated from K to R. In VH3-33, VH3-30, VH3-7, and VH3-48, this amino acid was mutated from R to K. In VH3-32, this position was mutated from K to R. The purpose of making these mutations was to enhance the diversity of CDRH3 presentation in the library. For example, in naturally occurring VH3-23 sequences, about 90% have K at position 94, while about 10% have position R. By making these changes the diversity of the CDRH3 presentation is increased, as is the overall diversity of the library.

Amplification was performed using the 424 vectors as a template. For the K94R mutation, the vectors containing the sequence DTAVYYCAK (VH3-23) were amplified with a PCR primer that changed the K to a R and added 5' tail for homologous recombination with the VH3-48, VH3-33, VH-30, and VH3-7. The "T" base in 3-48 does not change the amino acid encoded and thus the same primer with a T::C mismatch still allows homologous recombination into the 3-48 chassis.

Furthermore, the amplification products from the 424 vectors (produced as described above) containing the DTAVYYCAR sequence can be homologously recombined into the VH3-23 (CAR) vector, changing R to K in this framework and thus further increasing the diversity of CDRH3 presentation in this chassis.

```
                    240                                                294
VH3-48  (240)       TCTGCAAATGAACAGCCTGAGAGCTGAGGACACGGCGGTGTACTACTGCGCCAGA
VH3-33/30(240)      TCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCGGTGTACTACTGCGCCAGA
VH3-7  (240)        TCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCGGTGTACTACTGCGCCAGA
VH3-23  (240)       TCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCGGTGTACTACTGCGCCAAG
```

Example 11: VK Library Construction

This example describes the construction of a VK library of the invention. The exemplary VK library described herein corresponds to the VKCDR3 library of about 10^5 complexity, described in Example 6.2. As described in Example 6, and throughout the application, other VK libraries are within the scope of the invention, as are Vλ libraries.

Ten VK chassis were synthesized (Table 11), which did not contain VKCDR3, but instead had two SfiI restriction sites in the place of VKCDR3, as for the heavy chain vectors. The kappa constant region followed the SfiI restriction sites. Figure 8 shows a schematic structure of a light chain vector, prior to recombination with a CDRL3.

Ten VKCDR oligonucleotide libraries were then synthesized, as described in Example 6.2, using degenerate oligonucleotides (Table 33). The oligonucleotides were then PCR amplified, as separate pools, to make them double stranded and to add additional nucleotides required for efficient homologous recombination with the gapped (by SfiI) vector containing the VK chassis and constant region sequences. The VKCDR3 pools in this embodiment of the invention represented lengths 8, 9, and 10 amino acids, which were mixed post-PCR at a ratio 1:8:1. The pools were then cloned into the respective SfiI gapped VK chassis via homologous recombination, as described for the CDRH3 regions, set forth above. A schematic diagram of a CDRL3 integrated into a light chain vector and the accompanying sequence are provided in Figure 9.

A kappa light chain library pool was then produced, based on the approximate representation of the VK family members found in the circulating pool of B cells. The 10 kappa variable regions used and the relative frequency in the final library pool are shown in Table 44.

Table 44. Occurrence of VK Chassis in Data Sets Used to Design Library, Expected (Designed) Library, and Actual (Observed) Library

Chassis	Relative Occurrence in Data Sets (1)	Expected (2)	Observed (3)
VK1-5	8.6	7.1	5.8
VK1-12	4.0	3.6	3.5
VK1-27	3.3	3.6	8.1
VK1-33	5.3	7.1	3.5
VK1-39	18.5	21.4	17.4
VK2-28	7.7	7.1	5.8
VK3-11	10.9	10.7	20.9
VK3-15	6.6	7.1	4.7
VK3-20	24.5	21.4	18.6
VK4-1	10.4	10.7	11.6

(1) As indicated in Example 3, these 10 chassis account for about 80% of the occurrences in the entire data set of VK sequences examined.

(2) Rounded off ratios from the data in column 2, then normalized for actual experimental set up. The relative rounded ratios are 6 for VK1-39 and VK3-20, 3 for VK3-11 and VK4-1, 2 for VK-15, VK1-33, VK2-28 and VK3-15, and 1 for VK1-12 and VK1-27.

(3) Chassis usage in set of 86 sequences obtained from library; see also Figure 22.

Example 12: Characterization of Exemplary Libraries

This example shows the characteristics of exemplary libraries of the invention, constructed according to the methods described herein.

Example 12.1. Characterization of the Heavy Chains

To characterize the product of the split pool synthesis, ten of the 424 vectors containing the [Tail]-[N1]-[DH]-[N2]-[H3-JH] product were selected at random and transformed into *E. coli.* The split pool product had a theoretical diversity of about 1.1×10^6 (*i.e.,* 278 × 141 × 28). Ninety-six colonies were selected from the transformation and forward and reverse sequences were generated for each clone. Of the 96 sequencing reactions, 90 yielded sequences from which the CDRH3 region could be identified, and about 70% of these sequences matched a designed sequence in the library. The length distribution of the sequenced CDRH3 segments from the ten vectors, as compared to the theoretical distribution (based on design), is provided in Figure 10. The length distribution of the individual DH, N2, and H3-JH segments obtained from the ten vectors are shown in Figures 11-13.

Once the length distribution of the CDRH3 components of the library that were contained in the vector matched design were verified, the CDRH3 domains and heavy chain family representation in yeast that had been transformed according to the process described in Example 10.4 were characterized. Over 500 single-pass sequences were obtained. Of these, 531 yielded enough sequence information to identify the heavy chain chassis and 291 yielded enough sequence information to characterize the CDRH3. These CDRH3 domains have been integrated with the heavy chain chassis and constant region, according to the homologous recombination processes described herein. The length distribution of the CDRH3 domains from 291 sequences, compared to the theoretical length distribution, is shown in Figure 14. The mean theoretical length was 14.4 ± 4 amino acids, while the average observed length was 14.3 ± 3 amino acids. The observed length of each portion of the CDRH3, as compared to theoretical, is presented in Figures 15-18. Figure 19 depicts the familial origin of the JH segments identified in the 291 sequences, and Figure 20 shows the representation of 16 of the chassis of the library. The VH3-15 chassis was not represented amongst these sequences. This was corrected later by introducing yeast transformants containing the VH3-15 chassis, with CDRH3 diversity, into the library at the desired composition.

Example 12.2. Characterization of the Light Chains

The length distribution of the CDRL3 components, from the VKCDR3 library described in Example 6.2, were determined after yeast transformation via the methods described in Example 10.4. A comparison of the CDRL3 length from 86 sequences of the library to the human sequences and designed sequences is provided in Figure 21. Figure 22 shows the representation of the light chain chassis from amongst the 86 sequences selected from the library. About 91% of the CDRL3 sequences were exact matches to the design, and about 9% differed by a single amino acid.

Example 13: Characterization of the Composition of the Designed CDRH3 Libraries

This example presents data on the composition of the CDRH3 domains of exemplary libraries, and a comparison to other libraries of the art. More specifically, this example presents an analysis of the occurrence of the 400 possible amino acid pairs (20 amino acids × 20 amino acids) occurring in the CDRH3 domains of the libraries. The prevalence of these pairs is computed by examination of the nearest neighbor (i – i+1; designated IP1), next nearest neighbor (i – i+2; designated IP2), and next-next nearest neighbor (i – i+3; designated IP3) of the i residue in CDRH3. Libraries previously known in the art (*e.g.,* Knappik *et al.,* J. Mol. Biol., 2000, 296: 57; Sidhu *et al.,* J. Mol. Biol., 2004, 338: 299; and Lee *et al.*, J. Mol. Biol. 2004, 340: 1073, each of which is incorporated by reference in its entirety) have only considered the occurrence of the 20 amino acids at *individual* positions within CDRH3, while maintaining the same composition across the *center* of CDRH3, and not the pair-wise occurrences considered herein. In fact, according to Sidhu *et al.* (J. Mol. Biol., 2004, 338: 299, incorporated by reference in its entirety), "[i]n CDR-H3, there was some bias towards certain residue types, but all 20 natural amino acid residues occurred to a significant extent, and there was very little position-specific bias within the central portion of the loop". Thus, the present invention represents the first recognition that, surprisingly, a position-specific bias does exist within the central portion of the CDRH3 loop, when the occurrences of amino acid pairs recited above are considered. This example shows that the libraries described herein more faithfully reproduce the occurrence of these pairs as found in human sequences, in comparison to other libraries of the art. The composition of the libraries described herein may thus be considered more "human" than other libraries of the art.

To examine the pair-wise composition of CDRH3 domains, a portion of CDRH3 beginning at position 95 was chosen. For the purposes of comparison with data presented in Knappik *et al.* and Lee *et al.*, the last five residues in each of the analyzed CDRH3s were ignored. Thus, for the purposes of this analysis, both members of the pair i – i + X (X=1 to 3) must fall within the region starting at position 95 and ending at (but including) the sixth residue from the C-terminus of the CDRH3. The analyzed portion is termed the "central loop" (see Definitions).

To estimate pair distributions in representative libraries of the invention, a sampling approach was used. A number of sequences were generated by choosing randomly and, in turn, one of the 424 tail plus N1 segments, one of the 278 DH segments, one of the 141 N2 segments and one of the 28 JH segments (the latter truncated to include only the 95 to 102 Kabat CDRH3). The process was repeated 10,000 times to generate a sample of 10,000 sequences. By choosing a different seed for the random number generation, an independent sample of another 10,000 sequences was also generated and the results for pair distributions were observed to be nearly the same. For the calculations presented herein, a third and much larger sample of 50,000 sequences was used. A similar approach was used for the alternative library embodiment (N1-141), whereby the first segment was selected from 1068 tail+N1 segments (resulting after eliminating redundant sequences from 2 times 4 times 141 or 1128 possible combinations).

The pair-wise composition of Knappik *et al.* was determined based on the percent occurrences presented in Figure 7a of Knappik *et al.* (p.71). The relevant data are reproduced below, in Table 45.

Table 45. Composition of CDRH3 positions 95-100s (corresponding to positions 95-99B of the libraries of the current invention) of CDRH3 of Knappik *et al.* (from Figure

7a of Knappik *et al.*)

Amino Acid	Planned (%)	Found (%)
A	4.1	3.0
C	1.0	1.0
D	4.1	4.2
E	4.1	2.3
F	4.1	4.9
G	15.0	10.8
H	4.1	4.6
I	4.1	4.5
K	4.1	2.9
L	4.1	6.6
M	4.1	3.3
N	4.1	4.5
P	4.1	4.8
Q	4.1	2.9
R	4.1	4.1
S	4.1	5.6
T	4.1	4.5
V	4.1	3.7
W	4.1	2.0
Y	15.0	19.8

The pair-wise composition of Lee *et al.* was determined based on the libraries depicted in Table 5 of Lee *et al.*, where the positions corresponding to those CDRH3 regions analyzed from the current invention and from Knappik *et al.* are composed of an "XYZ" codon in Lee *et al.* The XYZ codon of Lee *et al.* is a degenerate codon with the following base compositions:

position 1 (X): 19% A, 17% C, 38% G, and 26% T;
position 2 (Y): 34% A, 18% C, 31% G, and 17% T; and
position 3 (Z): 24% G and 76% T.

When the approximately 2% of codons encoding stop codons are excluded (these do not occur in functionally expressed human CDRH3 sequences), and the percentages are re-normalized to 100%, the following amino acid representation can be deduced from the composition of the XYZ codon of Lee *et al.* (Table 46).

Table 46. Composition of CDRH3 of Lee *et al.*, Based on the Composition of the Degenerate XYZ Codon.

Type	Percent	Type	Percent
A	6.99%	M	0.79%
C	6.26%	N	5.02%
D	10.03%	P	3.13%
E	3.17%	Q	1.42%
F	3.43%	R	6.83%
G	12.04%	S	9.35%
H	4.49%	T	3.49%
I	2.51%	V	6.60%
K	1.58%	W	1.98%
L	4.04%	Y	6.86%

The occurrences of each of the 400 amino acid pairs, in each of the IP1, IP2, and IP3 configurations, can be computed for Knappik *et al.* and Lee *et al.* by multiplying together the individual amino acid compositions. For example, for Knappik *et al.,* the occurrence of YS pairs in the library is calculated by multiplying 15% by 4.1%, to yield 6.1%; note that the occurrence of SY pairs would be the same. Similarly, for the XYZ codon-based libraries of Lee *et al.,* the occurrence of YS pairs would be 6.86% (Y) multiplied by 9.35% (S), to give 6.4%; the same, again, for SY.

For the human CDRH3 sequences, the calculation is performed by ignoring the last five amino acids in the Kabat definition. By ignoring the C-terminal 5 amino acids of the human CDRH3, these sequences may be compared to those of Lee *et al.*, based on the XYZ codons. While Lee *et al.* also present libraries with “NNK” and “NNS” codons, the pair-wise compositions of these libraries are even further away from human CDRH3 pair-wise composition. The XYZ codon was designed by Lee *et al.* to replicate, to some extent, the individual amino acid type biases observed in CDRH3.

An identical approach was used for the libraries of the invention, after using the methods described above to produce sample sequences. While it is possible to perform these calculations with all sequences in the library, independent random samples of 10,000 to 20,000 members gave indistinguishable results. The numbers reported herein were thus generated from samples of 50,000 members.

Three tables were generated for IP1, IP2 and IP3, respectively (Tables 47, 48, and 49). Out of the 400 pairs, a selection from amongst the 20 most frequently occurring is included in the tables. The sample of about 1,000 human sequences (Lee *et al.,* 2006) is denoted as “Preimmune,” a sample of about 2,500 sequences (Jackson *et al.,* 2007) is denoted as “Humabs,” and the more affinity matured subset of the latter, which excludes all of the Preimmune set, is denoted as "Matured." Synthetic libraries in the art are denoted as HuCAL (Knappik, *et al.,* 2000) and XYZ (Lee *et al.,* e 2004). Two representative libraries of the invention are included: LUA-59 includes 59 N1 segments, 278 DH segments, 141 N2 segments, and 28 H3-JH segments (see Examples, above). LUA-141 includes 141 N1 segments, 278 DH segments, 141 N2 segments, and 28 H3-JH segments (see Examples, above). Redundancies created by combination of the N1 and tail sequences were removed from the dataset in each respective library. In certain embodiments, the invention may be defined based on the percent occurrence of any of the 400 amino acid pairs, particularly those in Tables 47-49. In certain embodiments, the invention may be defined based on at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more of these pairs. In certain embodiments of the invention, the percent occurrence of certain pairs of amino acids may fall within ranges indicated by "LUA-" (lower boundary) and "LUA+" (higher boundary), in the following tables. In some embodiments of the invention, the lower boundary for the percent occurrence of any amino acid pairs may be about 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, 4.75, and 5. In some embodiments of the invention, the higher boundary for the percent occurrence of any amino acid pairs may be about 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, 4.75, 5, 5.25, 5.5, 5.75, 6, 6.25, 6.5, 6.75, 7, 7.25, 7.5, 7.75, and 8. According to the present invention, any of the lower boundaries recited may be combined with any of the higher boundaries recited, to establish ranges, and vice-versa.

Table 47. Percent Occurrence of i – i+1 (IP1) Amino Acid Pairs in Human Sequences, Exemplary Libraries of the Invention, and the Libraries of Knappik *et al.* and Lee *et al.*

Pairs	Preimmune	Humabs	Matured	LUA-59	LUA-141	HuCAL	XYZ	LUA-	LUA+	Range	HuCAL	XYZ
YY	**5.87**	**4.44**	**3.27**	**5.83**	**5.93**	**2.25**	**0.47**	**2.50**	**6.50**	**4.00**	**0**	**0**
SG	**3.54**	**3.41**	**3.26**	**3.90**	**3.72**	**0.61**	**1.13**	**2.50**	**4.50**	**2.00**	**0**	**0**
SS	**3.35**	**2.65**	**2.26**	**2.82**	**3.08**	**0.16**	**0.88**	**2.00**	**4.00**	**2.00**	**0**	**0**
GS	**2.59**	**2.37**	**2.20**	**3.82**	**3.52**	**0.61**	**1.13**	**1.50**	**4.00**	**2.50**	**0**	**0**
GY	**2.55**	**2.34**	**2.12**	**3.15**	**2.56**	**2.25**	**0.83**	**2.00**	**3.50**	**1.50**	**1**	**0**
GG	**2.19**	**2.28**	**2.41**	**6.78**	**3.51**	**2.25**	**1.45**	**2.00**	**7.00**	**5.00**	**1**	**0**
YS	**1.45**	**1.30**	**1.23**	**1.40**	**1.52**	**0.61**	**0.64**	**0.75**	**2.00**	**1.25**	**0**	**0**
YG	**1.35**	**1.21**	**1.10**	**1.64**	**1.69**	**2.25**	**0.83**	**0.75**	**2.00**	**1.25**	**0**	**1**
SY	**1.31**	**1.07**	**0.90**	**1.65**	**1.77**	**0.61**	**0.64**	**0.75**	**2.00**	**1.25**	**0**	**0**
YD	1.67	1.40	1.17	0.88	0.90	0.61	0.69	0.75	2.25	1.50	0	0
DS	1.53	1.31	1.16	1.20	1.46	0.16	0.94	0.75	2.00	1.25	0	1
DY	1.40	1.23	1.11	0.34	0.48	0.61	0.69	0.25	2.00	1.75	1	1
VV	1.37	0.94	0.64	2.30	2.30	0.16	0.44	0.50	2.50	2.00	0	0
GD	1.20	1.21	1.25	0.49	0.44	0.61	1.21	0.25	1.75	1.50	1	1
AA	1.16	0.93	0.75	1.27	1.46	0.16	0.49	0.60	1.50	0.90	0	0
RG	1.08	1.26	1.38	1.69	1.38	0.61	0.82	1.00	2.00	1.00	0	0
VA	0.91	0.66	0.46	0.36	0.35	0.16	0.46	0.25	1.00	0.75	0	1
GV	0.84	0.89	0.95	2.87	2.16	0.61	0.79	0.80	3.00	2.20	0	0
CS	0.82	0.55	0.38	0.79	0.80	0.04	0.59	0.50	1.00	0.50	0	1
GR	0.74	0.90	1.00	1.01	0.79	0.61	0.82	0.70	1.25	0.55	0	1

The pairs in bold comprise about 19% to about 24% of occurrences (among the possible 400 pairs) for the Preimmune (Lee, *et al.*, 2006), Humabs (Jackson, *et al.*, 2007) and matured (Jackson minus Lee) sets. They account for about 27% to about 31% of the occurrences in the LUA libraries, but only about 12% in the HuCAL library and about 8% in the "XYZ" library. This is a reflection of the fact that pair-wise biases do exist in the human and LUA libraries, but not in the others. The last 2 columns indicate whether the corresponding pair-wise compositions fall within the LUA- and LUA+ boundaries: 0 if outside, 1 if within.

Table 48. Percent Occurrence of i – i+2 (IP2) Amino Acid Pairs in Human Sequences, Exemplary Libraries of the Invention, and the Libraries of Knappik *et al.* and Lee *et al.*

Pairs	Preimmune	Humabs	Matured	LUA-59	LUA-141	HuCAL	XYZ	LUA-	LUA+	Range	HuCAL	XYZ
YY	**3.57**	**2.59**	**1.78**	**2.99**	**3.11**	**2.25**	**0.47**	**2.5**	**4.5**	**2**	**0**	**0**
GY	**3.34**	**2.91**	**2.56**	**4.96**	**3.78**	**2.25**	**0.83**	**2.5**	**5.5**	**3**	**0**	**0**
SY	**2.94**	**2.41**	**2.01**	**3.03**	**3.42**	**0.61**	**0.64**	**2**	**4**	**2**	**0**	**0**
YS	**2.88**	**2.34**	**1.95**	**3.24**	**3.32**	**0.61**	**0.64**	**1.75**	**3.75**	**2**	**0**	**0**
SG	**2.60**	**2.29**	**2.05**	**2.84**	**2.96**	**0.61**	**1.13**	**2**	**3.5**	**1.5**	**0**	**0**
SS	**2.27**	**2.01**	**1.84**	**2.30**	**2.50**	**0.16**	**0.88**	**1.5**	**3**	**1.5**	**0**	**0**
GS	**2.16**	**2.12**	**2.10**	**2.96**	**2.32**	**0.61**	**1.13**	**1.5**	**3**	**1.5**	**0**	**0**
GG	**1.92**	**2.25**	**2.44**	**6.23**	**3.68**	**2.25**	**1.45**	**1.5**	**7**	**5.5**	**1**	**0**
YG	**1.17**	**1.14**	**1.15**	**1.39**	**1.47**	**2.25**	**0.83**	**1**	**2**	**1**	**0**	**0**
DS	2.03	1.67	1.40	1.21	1.48	0.16	0.94	1	2.5	1.5	0	0
YD	1.71	1.39	1.11	0.89	0.92	0.61	0.69	0.75	1.75	1	0	0
VG	1.35	1.17	1.01	1.75	1.54	0.61	0.79	1	2	1	0	0
DY	1.06	1.02	0.99	0.23	0.40	0.61	0.69	0.2	1.2	1	1	1
WG	1.06	0.76	0.53	0.85	0.91	0.61	0.24	0.75	1.25	0.5	0	0
RY	0.98	1.00	0.96	0.70	0.91	0.61	0.47	0.6	1	0.4	1	0
GC	0.97	0.75	0.64	0.94	0.81	0.15	0.75	0.5	1	0.5	0	1
DG	0.95	1.05	1.08	1.78	1.05	0.61	1.21	0.75	2	1.25	0	1
GD	0.94	0.88	0.86	0.47	0.36	0.61	1.21	0.25	1	0.75	1	0
VV	0.94	0.59	0.35	0.95	0.90	0.16	0.44	0.5	1	0.5	0	0
AA	0.90	0.73	0.59	0.72	0.74	0.16	0.49	0.5	1	0.5	0	0

The pairs in bold comprise about 18% to about 23% of occurrences (among the possible 400 pairs) for the Preimmune (Lee, *et al.*, 2006), Humabs (Jackson, *et al.*, 2007) and matured (Jackson minus Lee) sets. They account for about 27% to about 30% of the occurrences in the LUA libraries, but only about 12% in the HuCAL library and about 8% in the "XYZ" library. Because of the nature of the construction of the central loops in the HuCAL and XYZ libraries, these numbers are the same for the IP1, IP2, and IP3 pairs. The last 2 columns indicate whether the corresponding pair-wise compositions fall within the LUA- and LUA+ boundaries: 0 if outside, 1 if within.

Table 49. Percent Occurrence of i – i+3 (IP3) Amino Acid Pairs in Human Sequences, Exemplary Libraries of the Invention, and the Libraries of Knappik *et al.* and Lee *et al.*

Pairs	Preimmune	Humabs	Matured	LUA-59	LUA-141	HuCAL	XYZ	LUA-	LUA+	Range	HuCAL	XYZ
GY	**3.55**	**2.85**	**2.32**	**5.80**	**4.42**	**2.25**	**0.83**	**2.5**	**6.5**	**4**	**0**	**0**
SY	**3.38**	**3.01**	**2.67**	**3.78**	**4.21**	**0.61**	**0.64**	**1**	**5**	**4**	**0**	**0**
YS	**3.18**	**2.56**	**2.05**	**3.20**	**3.33**	**0.61**	**0.64**	**2**	**4**	**2**	**0**	**0**
SS	**2.26**	**1.74**	**1.37**	**1.81**	**2.18**	**0.16**	**0.88**	**1**	**3**	**2**	**0**	**0**
GS	**2.23**	**2.13**	**2.00**	**4.60**	**3.33**	**0.61**	**1.13**	**2**	**5**	**3**	**0**	**0**
YG	**2.14**	**1.65**	**1.35**	**2.69**	**2.79**	**2.25**	**0.83**	**1.5**	**3**	**1.5**	**1**	**0**
YY	**1.86**	**1.48**	**1.12**	**1.18**	**1.27**	**2.25**	**0.47**	**0.75**	**2**	**1.25**	**0**	**0**
GG	**1.60**	**1.87**	**2.11**	**4.73**	**2.84**	**2.25**	**1.45**	**1.5**	**5**	**3.5**	**1**	**0**
SG	**0.90**	**1.04**	**1.12**	**0.93**	**1.25**	**0.61**	**1.13**	**0.75**	**1.5**	**0.75**	**0**	**1**
DG	2.01	1.94	1.84	2.51	2.03	0.61	1.21	1.5	3	1.5	0	0
DS	1.48	1.31	1.22	0.41	0.55	0.16	0.94	0.25	1.5	1.25	0	1
VA	1.18	0.83	0.55	1.48	1.46	0.16	0.46	0.5	2	1.5	0	0
AG	1.13	1.09	1.03	0.97	1.04	0.61	0.84	0.9	2	1.1	0	0
TY	1.05	0.90	0.76	1.01	1.16	0.61	0.24	0.75	1.75	1	0	0
PY	1.02	0.88	0.79	1.23	0.86	0.61	0.21	0.75	1.75	1	0	0
RS	1.02	0.88	0.77	0.38	0.55	0.16	0.64	0.25	1.25	1	0	1
RY	1.02	1.12	1.14	0.68	0.88	0.61	0.47	0.65	1.25	0.6	0	0
LY	1.01	0.88	0.75	0.69	0.76	0.61	0.28	0.65	1.25	0.6	0	0
DY	0.93	0.84	0.77	0.72	0.95	0.61	0.69	0.7	1.3	0.6	0	0
GC	0.90	0.62	0.48	0.86	0.68	0.15	0.75	0.5	1	0.5	0	1

The pairs in bold make up about 16 to about 21% of the occurrences (among the possible 400 pairs) for the Preimmune (Lee, *et al.,* 2006), Humabs (Jackson, *et al.,* 2007) and matured (Jackson minus Lee) sets. They account for 26 to 29% of the occurrences in the LUA libraris, but only about 12% in the HuCAL library and about 8% for the "XYZ" library. Because of the nature of the construction of the central loops in the HuCAL and XYZ libraries, these numbers are the same for the IP1, IP2, and IP3 pairs. The last 2 columns indicate whether the corresponding pair-wise compositions fall within the LUA- and LUA+ boundaries: 0 if outside, 1 if within.

The analysis provided in this example demonstrates that the composition of the libraries of the present invention more closely mimics the composition of human sequences than other libraries known in the art. Synthetic libraries of the art do not intrinsically reproduce the composition of the "central loop" portion actual human CDRH3 sequences at the level of pair percentages. The libraries of the invention have a more complex pair-wise composition that closely reproduces that observed in actual human CDRH3 sequences. The exact degree of this reproduction versus a target set of actual human CDRH3 sequences may be optimized, for example, by varying the compositions of the segments used to design the CDRH3 libraries. Moreover, it is also possible to utilize these metrics to computationally design libraries that exactly mimic the pair-wise compositional prevalence found in human sequences.

Example 14: Information Content of Exemplary Libraries

One way to quantify the observation that certain libraries, or collection of sequences, may be intrinsically more complex or "less random" than others is to apply information theory (Shannon, Bell Sys. Tech. J., 1984, 27: 379; Martin *et al.,* Bioinformatics, 2005, 21: 4116; Weiss *et al.,* J. Theor. Biol., 2000, 206: 379, each incorporated by reference in its entirety). For example, a metric can be devised to quantify the fact that a position with a fixed amino acid represents less "randomness" than a position where all 20 amino acids may occur with equal probability. Intermediate situations should lead, in turn, to intermediate values of such a metric. According to information theory this metric can be represented by the formula:

$$I = \sum_{i=1}^{N} f_i \log_2 f_i$$

Here, f_i is the normalized frequency of occurrence of i, which may be an amino acid type (in which case N would be equal to 20). When all f_i are zero except for one, the value of I is zero. In any other case the value of I would be smaller, *i.e.,* negative, and the lowest value is achieved when all f_i values are the same and equal to N. For the amino acid case, N is 20, and the resulting value of I would be -4.322. Because I is defined with base 2 logarithms, the units of I are bits.

The I value for the HuCAL and XYZ libraries at the single position level may be derived from Tables 45 and 46, respectively, and are equal to -4.08 and -4.06. The corresponding single residue frequency occurrences in the non-limiting exemplary libraries of the invention and the sets of human sequences previously introduced, taken within the "central loop" as defined above, are provided in Table 50.

Table 50. Amino Acid Type Frequencies in Central Loop

Type	Preimmune	Humabs	Matured	LUA-59	LUA-141
A	5.46	5.51	5.39	5.71	6.06
C	1.88	1.46	1.22	1.33	1.34
D	7.70	7.51	7.38	4.76	5.23
E	2.40	2.90	3.28	3.99	4.68
F	2.29	2.60	2.81	1.76	2.17
G	14.86	15.42	15.82	24.90	18.85
H	1.46	1.79	2.01	0.20	0.67
I	3.71	3.26	2.99	3.99	4.34
K	1.06	1.27	1.44	0.21	0.67
L	4.48	4.84	5.16	4.12	4.54
M	1.18	1.03	0.93	0.94	1.03
N	1.81	2.43	2.84	0.41	0.65
P	4.12	4.10	4.13	5.68	3.96
Q	1.60	1.77	1.95	0.21	0.68
R	5.05	5.90	6.41	3.35	4.11
S	12.61	11.83	11.37	11.18	12.77
T	4.59	5.11	5.47	4.36	4.95
V	6.21	5.55	5.12	8.13	7.67
W	2.79	2.91	3.07	1.57	1.98
Y	14.74	12.81	11.24	13.20	13.63

The information content of these sets, computed by the formula given above, would then be -3.88, -3.93, -3.96, -3.56, and -3.75, for the preimmune, human, matured, LUA-59 and LUA-141 sets, respectively. As the frequencies deviate more from completely uniform (5% for each of the 20), then numbers tend to be larger, or less negative.

The identical approach can be used to analyze pair compositions, or frequencies, by calculating the sum in the formula above over the 20x20 or 400 values of the frequencies for each of the pairs. It can be shown that any pair frequency made up of the simple product of two singleton frequency sets is equal to the sum of the individual singleton I values. If the two singleton frequency sets are the same or approximately so, this means that I (independent pairs) = 2 * I (singles). It is thus possible to define a special case of the mutual information, MI, for a general set of pair frequencies as MI (pair) = I(pair) – 2 * I (singles) to measure the amount of information gained by the structure of the pair frequencies themselves (compare to the standard definitions in Martin *et al.,* 2005, for example, after considering that I (X) = -H(X) in their notation). When there is no such structure, the value of MI is simply zero.

Values of MI computed from the pair distributions discussed above (over the entire set of 400 values) are given in Table 51.

Table 51. Mutual Information Within Central Loop of CDRH3

Library or Set	**i - i+1**	**i - i+2**	**i - i+3**
Preimmune	0.226	0.192	0.163
Humabs	0.153	0.128	0.111
Matured	0.124	0.107	0.100
LUA-59	0.422	0.327	0.278
LUA-141	0.376	0.305	0.277
HuCAL	0.000	0.000	0.000
XYZ	0.000	0.000	0.000

It is notable that the MI values decrease within sets of human sequences as those sequences undergo further somatic mutation, a process that over many independent sequences is essentially random. It is also worth noting that the MI values decrease as the pairs being considered sit further and further apart, and this is the case for both sets of human sequences, and exemplary libraries of the invention. In both cases, as the two amino acids in a pair become further separated the odds of their straddling an actual segment (V, D, J plus V-D or D-J insertions) increase, and their pair frequencies become closer to a simple product of singleton frequencies.

Table 52 contains sequence information on certain immunoglobulin gene segments cited in the application. These sequences are non-limiting, and it is recognized that allelic variants exist and encompassed by the present invention. Accordingly, the methods present herein can be utilized with mutants of these sequences.

Table 52. Sequence Information for Certain Immunoglobulin Gene Segments Cited Herein

SEQ ID NO:	Sequence	Peptide or Nucleotide Sequence	Observations
	IGHV1-3	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYAMHWVRQAPGQRLEWMGWINAGNGNTKYSQKFQGRVTITRDTSASTAYMELSSLRSEDTAVYYCAR	
	IGHV1-8_v1	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYDINWVRQATGQGLEWMGWMNPNSGNTGYAQKFQGRVTMTR**NTS**ISTAYMELSSLRSEDTAVYYCAR	
	IGHV1-8_v2	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYDINWVRQATGQGLEWMGWMNPNSGNTGYAQKFQGRVTMTR**DTS**ISTAYMELSSLRSEDTAVYYCAR	N to D mutation avoids NTS potential glycosylation site in the original germline sequence (v1 above). XTS, where X is not N, and NTZ, where Z is not S or T are also options. NPS is yet another option that is much less likely to be N-linked glycosylated.
	IGHV1-24	QVQLVQSGAEVKKPGASVKVSCKVSGYTLTELSMHWVRQAPGKGLEWMGGFDPEDGETIYAQKFQGRVTMTEDTSTDTAYMELSSLRSEDTAVYYCAT	
	IGHV1-45	QMQLVQSGAEVKKTGSSVKVSCKASGYTFTYRYLHWVRQAPGQALEWMGWITPFNGNTNYAQKFQDRVTITRDRSMSTAYMELSSLRSEDTAMYYCAR	

	IGHV1-58	QMQLVQSGPEVKKPGTSVKVSCKASGFTFTSSAVQWVRQARGQRLEWIGWIVVGSGNTNYAQKFQERVTITRDMSTSTAYMELSSLRSEDTAVYYCAA	
	IGHV2-5	QITLKESGPTLVKPTQTLTLTCTFSGFSLSTSGVGVGWIRQPPGKALEWLALIYWDDDKRYSPSLKSRLTITKDTSKNQVVLTMTNMDPVDTATYYCAHR	
	IGHV2-26	QVTLKESGPVLVKPTETLTLTCTVSGFSLSNARMGVSWIRQPPGKALEWLAHIFSNDEKSYSTSLKSRLTISKDTSKSQVVLTMTNMDPVDTATYYCARI	
	IGHV2-70_v1	RVTLRESGPALVKPTQTLTLTCTFSGFSLSTSGM**C**VSWIRQPPGKALEWLARIDWDDDKYYSTSLKTRLTISKDTSKNQVVLTMTNMDPVDTATYYCARI	
	IGHV2-70_v2	RVTLRESGPALVKPTQTLTLTCTFSGFSLSTSGM**G**VSWIRQPPGKALEWLARIDWDDDKYYSTSLKTRLTISKDTSKNQVVLTMTNMDPVDTATYYCARI	C to G mutation avoids unpaired Cys in v1 above. G was chosen by analogy to other germline sequences, but other amino acid types, R, S, T, as non-limiting examples, are possible.
	IGHV3-9	EVQLVESGGGLVQPGRSLRLSCAASGFTFDDYAMHWVRQAPGKGLEWVSGISWNSGSIGYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTALYYCAKD	
	IGHV3-11	QVQLVESGGGLVKPGGSLRLSCAASGFTFSDYYMSWIRQAPGKGLEWVSYISSSGSTIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR	
	IGHV3-13	EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYDMHWVRQATGKGLEWVSAIGTAGDTYYPGSVKGRFTISRENAKNSLYLQMNSLRAGDTAVYYCAR	
	IGHV3-20	EVQLVESGGGVVRPGGSLRLSCAASGFTFDDYGMSWVRQAPGKGLEWVSGINWNGGSTGYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTALYHCAR	
	IGHV3-21	EVQLVESGGGLVKPGGSLRLSCAASGFTFSSYSMNWVRQAPGKGLEWVSSISSSSSYIYYADSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR	

	IGHV3-43	EVQLVESGGVVVQPGGSLRLSCAASGFTFDDYTMHWVRQAPGKGLEWVSLISWDGGSTYYADSVKGRFTISRDNSKNSLYLQMNSLRTEDTALYYCAKD	
	IGHV3-49	EVQLVESGGGLVQPGRSLRLSCTASGFTFGDYAMSWVRQAPGKGLEWVGFIRSKAYGGTTEYAASVKGRFTISRDDSKSIAYLQMNSLKTEDTAVYYCTR	
	IGHV3-53	EVQLVESGGGLIQPGGSLRLSCAASGFTVSSNYMSWVRQAPGKGLEWVSVIYSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR	
	IGHV3-64	EVQLVESGGGLVQPGGSLRLSCSASGFTFSSYAMHWVRQAPGKGLEYVSAISSNGGSTYYADSVKGRFTISRDNSKNTLYLQMSSLRAEDTAVYYCVK	
	IGHV3-66	EVQLVESGGGLVQPGGSLRLSCAASGFTVSSNYMSWVRQAPGKGLEWVSVIYSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAR	
	IGHV3-72	EVQLVESGGGLVQPGGSLRLSCAASGFTFSDHYMDWVRQAPGKGLEWVGRTRNKANSYTTEYAASVKGRFTISRDDSKNSLYLQMNSLKTEDTAVYYCAR	
	IGHV3-73	EVQLVESGGGLVQPGGSLKLSCAASGFTFSGSAMHWVRQASGKGLEWVGRIRSKANSYATAYAASVKGRFTISRDDSKNTAYLQMNSLKTEDTAVYYCTR	
	IGHV3-74	EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMHWVRQAPGKGLVWVSRINSDGSSTSYADSVKGRFTISRDNAKNTLYLQMNSLRAEDTAVYYCAR	
	IGHV4-4v1	QVQLQESGPGLVKPSGTLSLTCAVSGGSISSSNWWSWVRQPPGKGLEWIGEIYHSGSTNYNPSLKSRVTISVDKSKNQFSLKLSSVTAADTAVYYCAR	Contains CDRH1 with size 6 (Kabat definition); canonical structure H1-2. Sequence corresponds to allele *02 of IGHV4-4.
	IGHV4-4v2	QVQLQESGPGLVKPSETLSLTCTVSGGSISSYYWSWIRQPAGKGLEWIGRIYTSGSTNYNPSLKSRVTMSVDTSKNQFSLKLSSVTAADTAVYYCAR	Contains CDRH1 with size 5 (Kabat definition); canonical structure H1-1. Sequence corresponds to allele *07 of

			IGHV4-4
	IGHV4-28	QVQLQESGPGLVKPSDTLSLTCAVSGYSISSSNWWGWIRQPPGKGLEWIGYIYYSGSTYYNPSLKSRVTMSVDTSKNQFSLKLSSVTAVDTAVYYCAR	
	IGHV6-1	QVQLQQSGPGLVKPSQTLSLTCAISGDSVSSNSAAWNWIRQSPSRGLEWLGRTYYRSKWYNDYAVSVKSRITINPDTSKNQFSLQLNSVTPEDTAVYYCAR	
	IGHV7-4-1	QVQLVQSGSELKKPGASVKVSCKASGYTFTSYAMNWVRQAPGQGLEWMGWINTNTGNPTYAQGFTGRFVFSLDTSVSTAYLQISSLKAEDTAVYYCAR	
	IGKV1-06	AIQMTQSPSSLSASVGDRVTITCRASQGIRNDLGWYQQKPGKAPKLLIYAASSLQSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCLQDYNYP	
	IGKV1-08_v1	AIRMTQSPSSFSASTGDRVTITCRASQGISSYLAWYQQKPGKAPKLLIYAASTLQSGVPSRFSGSGSGTDFTLTIS**C**LQSEDFATYYCQQYYSYP	
	IGKV1-08_v2	AIRMTQSPSSFSASTGDRVTITCRASQGISSYLAWYQQKPGKAPKLLIYAASTLQSGVPSRFSGSGSGTDFTLTIS**S**LQSEDFATYYCQQYYSYP	C to S mutation avoids unpaired Cys. in v1 above. S was chosen by analogy to other germline sequences, but amino acid types, N, R, S, as non-limiting examples, are also possible
	IGKV1-09	DIQLTQSPSFLSASVGDRVTITCRASQGISSYLAWYQQKPGKAPKLLIYAASTLQSGVPSRFSGSGSGTEFTLTISSLQPEDFATYYCQQLNSYP	
	IGKV1-13	AIQLTQSPSSLSASVGDRVTITCRASQGISSALAWYQQKPGKAPKLLIYDASSLESGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCQQFNSYP	
	IGKV1-16	DIQMTQSPSSLSASVGDRVTITCRASQGISNYLAWFQQKPGKAPKSLIYAASSLQSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCQQYNSYP	
	IGKV1-17	DIQMTQSPSSLSASVGDRVTITCRASQGIRNDLGWYQQKP	

		GKAPKRLIYAASSLQSGVPSRFSGSGSGTEFTLTISSLQPEDFATYYCLQHNSYP	
	IGKV1-37_v1	DIQLTQSPSSLSASVGDRVTITCRVSQGISSYLNWYRQKPGKVPKLLIYSASNLQSGVPSRFSGSGSGTDFTLTISSLQPEDVATYY**G**QRTYNAP	
	IGKV1-37_v2	DIQLTQSPSSLSASVGDRVTITCRVSQGISSYLNWYRQKPGKVPKLLIYSASNLQSGVPSRFSGSGSGTDFTLTISSLQPEDVATYY**C**QRTYNAP	Restores conserved Cys, missing in v1 above, just prior to CDRL3.
	IGKV1D-16	DIQMTQSPSSLSASVGDRVTITCRASQGISSWLAWYQQKPEKAPKSLIYAASSLQSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCQQYNSYP	
	IGKV1D-17	NIQMTQSPSAMSASVGDRVTITCRARQGISNYLAWFQQKPGKVPKHLIYAASSLQSGVPSRFSGSGSGTEFTLTISSLQPEDFATYYCLQHNSYP	
	IGKV1D-43	AIRMTQSPFSLSASVGDRVTITCWASQGISSYLAWYQQKPAKAPKLFIYYASSLQSGVPSRFSGSGSGTDYTLTISSLQPEDFATYYCQQYYSTP	
	IGKV1D-8_v1	VIWMTQSPSLLSASTGDRVTISCRMSQGISSYLAWYQQKPGKAPELLIYAASTLQSGVPSRFSGSGSGTDFTLTIS**C**LQSEDFATYYCQQYYSFP	
	IGKV1D-8_v2	VIWMTQSPSLLSASTGDRVTISCRMSQGISSYLAWYQQKPGKAPELLIYAASTLQSGVPSRFSGSGSGTDFTLTIS**S**LQSEDFATYYCQQYYSFP	C to S mutation avoids unpaired Cys. in v1 above. S was chosen by analogy to other germline sequences, but amino acid types, N, R, S, as non-limiting examples, are also possible
	IGKV2-24	DIVMTQTPLSSPVTLGQPASISCRSSQSLVHSDGNTYLSWLQQRPGQPPRLLIYKISNRFSGVPDRFSGSGAGTDFTLKISRVEAEDVGVYYCMQATQFP	
	IGKV2-29	DIVMTQTFLSLSVTRQQPASISCKSSQSLLHSDGVTYLYWYLQRPQQSPQLLTYEVSSRFSGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQGTHLP	
	IGKV2-30	DVVMTQSPLSLPVTLGQPASISCRSSQSLVYSDGNTYLNW	

		FQQRPGQSPRRLIYKVSNRDSGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQGTHWP	
	IGKV2-40	DIVMTQTPLSLPVTPGEPASISCRSSQSLLDSDDGNTYLDWYLQKPGQSPQLLIYTLSYRASGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQRIEFP	
	IGKV2D-26	EIVMTQTPLSLSITPGEQASMSCRSSQSLLHSDGYTYLYWFLQKARPVSTLLIYEVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDFGVYYCMQDAQD	
	IGKV2D-29	DIVMTQTPLSLSVTPGQPASISCKSSQSLLHSDGKTYLYWYLQKPGQPPQLLIYEVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQSIQLP	
	IGKV2D-30	DVVMTQSPLSLPVTLGQPASISCRSSQSLVYSDGNTYLNWFQQRPGQSPRRLIYKVSNWDSGVPDRFSGSGSGTDFTLKISRVEAEDVGVYYCMQGTHWP	
	IGKV3D-07	EIVMTQSPATLSLSPGERATLSCRASQSVSSSYLSWYQQKPGQAPRLLIYGASTRATGIPARFSGSGSGTDFTLTISSLQPEDFAVYYCQQDYNLP	
	IGKV3D-11	EIVLTQSPATLSLSPGERATLSCRASQGVSSYLAWYQQKPGQAPRLLIYDASNRATGIPARFSGSGPGTDFTLTISSLEPEDFAVYYCQQRSNWH	
	IGKV3D-20	EIVLTQSPATLSLSPGERATLSCGASQSVSSSYLAWYQQKPGLAPRLLIYDASSRATGIPDRFSGSGSGTDFTLTISRLEPEDFAVYYCQQYGSSP	
	IGKV5-2_v1	ETTLTQSPAFMSATPGDKV**NIS**CKASQDIDDDMNWYQQKPGEAAIFIIQEATTLVPGIPPRFSGSGYGTDFTLTINNIESEDAAYYFCLQHDNFP	
	IGKV5-2_v2	ETTLTQSPAFMSATPGDKV**TIS**CKASQDIDDDMNWYQQKPGEAAIFIIQEATTLVPGIPPRFSGSGYGTDFTLTINNIESEDAAYYFCLQHDNFP	N to D mutation avoids NIS potential glycosylation site in v1 above. XIS, where X is not N, and NIZ, where Z is not S or T are also options. NPS is yet another option that is much less likely to be N-linked glycosylated.

	IGKV6-21	EIVLTQSPDFQSVTPKEKVTITCRASQSIGSSLHWYQQKPDQSPKLLIKYASQSFSGVPSRFSGSGSGTDFTLTINSLEAEDAATYYCHQSSSLP	
	IGKV6D-21	EIVLTQSPDFQSVTPKEKVTITCRASQSIGSSLHWYQQKPDQSPKLLIKYASQSFSGVPSRFSGSGSGTDFTLTINSLEAEDAATYYCHQSSSLP	
	IGKV7-3	DIVLTQSPASLAVSPGQRATITCRASESVSFLGINLIHWYQQKPGQPPKLLIYQASNKDTGVPARFSGSGSGTDFTLTINPVEANDTANYYCLQSKNFP	
	IGλV1-36	QSVLTQPPSVSEAPRQRVTISCSGSSSNIGNNAVNWYQQLPGKAPKLLIYYDDLLPSGVSDRFSGSKSGTSASLAISGLQSEDEADYYCAAWDDSLNG	
	IGλV1-47	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNYVYWYQQLPGTAPKLLIYRNNQRPSGVPDRFSGSKSGTSASLAISGLRSEDEADYYCAAWDDSLSG	
	IGλV10-54	QAGLTQPPSVSKGLRQTATLTCTGNSNNVGNQGAAWLQQHQGHPPKLLSYRNNNRPSGISERLSASRSGNTASLTITGLQPEDEADYYCSAWDSSLSA	
	IGλV2-11_v1	QSALTQPRSVSGSPGQSVTISCTGTSSDVGGYNYVSWYQQHPGKAPKLMIYDVSKRPSGVPDRFSGSKSGNTASLTISGLQAEDEADYYC**C**SYAGSYTF	
	IGλV2-11_v2	QSALTQPRSVSGSPGQSVTISCTGTSSDVGGYNYVSWYQQHPGKAPKLMIYDVSKRPSGVPDRFSGSKSGNTASLTISGLQAEDEADYYC**S**SYAGSYTF	C to S mutation avoids unpaired Cys in v1 above. S was chosen by analogy to other germline sequences, but other amino acid types, such as Q, G, A, L, as non-limiting examples, are also possible
	IGλV2-18	QSALTQPPSVSGSPGQSVTISCTGTSSDVGSYNRVSWYQQPPGTAPKLMIYEVSNRPSGVPDRFSGSKSGNTASLTISGLQAEDEADYYCSLYTSSSTF	
	IGλV2-23_v1	QSALTQPASVSGSPGQSITISCTGTSSDVGSYNLVSWYQQ	

		HPGKAPKLMIYEGSKRPSGVSNRFSGSKSGNTASLTISGLQAEDEADYYC**C**SYAGSSTL	
	IGλV2-23_v2	QSALTQPASVSGSPGQSITISCTGTSSDVGSYNLVSWYQQHPGKAPKLMIYEGSKRPSGVSNRFSGSKSGNTASLTISGLQAEDEADYYC**S**SYAGSSTL	C to S mutation avoids unpaired Cys in v1 above. S was chosen by analogy to other germline sequences, but other amino acid types, such as Q, G, A, L, as non-limiting examples, are also possible
	IGλV2-8	QSALTQPPSASGSPGQSVTISCTGTSSDVGGYNYVSWYQQHPGKAPKLMIYEVSKRPSGVPDRFSGSKSGNTASLTVSGLQAEDEADYYCSSYAGSNNF	
	IGλV3-10	SYELTQPPSVSVSPGQTARITCSGDALPKKYAYWYQQKSGQAPVLVIYEDSKRPSGIPERFSGSSSGTMATLTISGAQVEDEADYYCYSTDSSGNH	
	IGλV3-12	SYELTQPHSVSVATAQMARITCGGNNIGSKAVHWYQQKPGQDPVLVIYSDSNRPSGIPERFSGSNPGNTTTLTISRIEAGDEADYYCQVWDSSSDH	
	IGλV3-16	SYELTQPPSVSVSLGQMARITCSGEALPKKYAYWYQQKPGQFPVLVIYKDSERPSGIPERFSGSSSGTIVTLTISGVQAEDEADYYCLSADSSGTY	
	IGλV3-25	SYELMQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTY	
	IGλV3-27	SYELTQPSSVSVSPGQTARITCSGDVLAKKYARWFQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGAQVEDEADYYCYSAADNN	
	IGλV3-9	SYELTQPLSVSVALGQTARITCGGNNIGSKNVHWYQQKPGQAPVLVIYRDSNRPSGIPERFSGSNSGNTATLTISRAQAGDEADYYCQVWDSSTA	
	IGλV4-3	LPVLTQPPSASALLGASIKLTCTLSSEHSTYTIEWYQQRPGRSPQYIMKVKSDGSHSKGDGIPDRFMGSSSGADRYLTFSNLQSDDEAEYHCGESHTIDGQVG	

	IGλV4-60	QPVLTQSSSASASLGSSVKLTCTLSSGHSSYIIAWHQQQPGKAPRYLMKLEGSGSYNKGSGVPDRFSGSSSGADRYLTISNLQLEDEADYYCETWDSNT	
	IGλV5-39	QPVLTQPTSLSASPGASARFTCTLRSGINVGTYRIYWYQQKPGSLPRYLLRYKSDSDKQQGSGVPSRFSGSKDASTNAGLLLISGLQSEDEADYYCAIWYSSTS	
	IGλV7-46	QAVVTQEPSLTVSPGGTVTLTCGSSTGAVTSGHYPYWFQQKPGQAPRTLIYDTSNKHSWTPARFSGSLLGGKAALTLSGAQPEDEAEYYCLLSYSGAR	
	IGλV8-61	QTVVTQEPSFSVSPGGTVTLTCGLSSGSVSTSYYPSWYQQTPGQAPRTLIYSTNTRSSGVPDRFSGSILGNKAALTITGAQADDESDYYCVLYMGSGI	
	IGλV9-49	QPVLTQPPSASASLGASVTLTCTLSSGYSNYKVDWYQQRPGKGPRFVMRVGTGGIVGSKGDGIPDRFSVLGSGLNRYLTIKNIQEEDESDYHCGADHGSGSNFV	
	IGHD1-1	GGTACAACTGGAACGAC	See (1) below.
	IGHD1-14	GGTATAACCGGAACCAC	
	IGHD1-20	GGTATAACTGGAACGAC	
	IGHD1-7	GGTATAACTGGAACTAC	
	IGHD2-21_v1	AGCATATTGTGGTGGTGA**T**TGCTATTCC	
	IGHD2-21_v2	AGCATATTGTGGTGGTGA**C**TGCTATTCC	Common allelic variant encoding a different amino acid sequence, compared to v1, in 2 of 3 forward reading frames.
	IGHD2-8	AGGATATTGTACTAATGGTGTATGCTATACC	
	IGHD3-16	GTATTATGATTACGTTTGGGGGAGTTATGCTTATACC	
	IGHD3-9	GTATTACGATATTTTGACTGGTTATTATAAC	
	IGHD4-23	TGACTACGGTGGTAACTCC	
	IGHD4-4/4-11	TGACTACAGTAACTAC	
	IGHD5-12	GTGGATATAGTGGCTACGATTAC	
	IGHD5-24	GTAGAGATGGCTACAATTAC	

	IGHD6-25	GGGTATAGCAGCGGCTAC	
	IGHD6-6	GAGTATAGCAGCTCGTCC	
	IGHD7-27	CTAACTGGGGA	

(1) Each of the IGHD nucleotide sequences can be read in three (3) forward reading frames, and, possibly, in 3 reverse reading frames. For example, the nucleotide sequence given for IGHD1-1, depending on how it inserts in full V-DJ rearrangement, may encode the full peptide sequences: GTTGT, VQLER and YNWND in the forward direction, and VVPVV, SFQLY and RSSCT in the reverse direction. Each of these sequences, in turn, could generate progressively deleted segments as explained in the Examples to produce suitable components for libraries of the invention.

Example 15: Selection of Antibodies from the Library

In this example, the selection of antibodies from a library of the invention (described in Examples 9-11 and other Examples) is demonstrated. These selections demonstrate that the libraries of the invention encode antibody proteins capable of binding to antigens. In one selection, antibodies specific for "Antigen X", a protein antigen, were isolated from the library using the methods described herein. Figure 24 shows binding curves for six clones specifically binding Antigen X, and their Kd values. This selection was performed using yeast with the heavy chain on a plasmid vector and the kappa light chain library integrated into the genome of the yeast.

In a separate selection, antibodies specific for a model antigen, hen egg white lysozyme (HEL) were isolated. Figure 25 shows the binding curves for 10 clones specifically binding HEL; each gave a Kd >500nM. This selection was performed using yeast with the heavy chain on a plasmid vector and the kappa light chain library on a plasmid vector. The sequences of the heavy and light chains were determined for clones isolated from the library and it was demonstrated that multiple clones were present. A portion of the FRM3s (underlined) and the entire CDRH3s from four clones are shown below (Table 53 and Table 54, the latter using the numbering system of the invention).

Table 53. Sequences of CDRH3, and a Portion of FRM3, from Four HEL Binders

Seq Name	SEQ ID NO:	FRM3 and CDRH3	Tail	N1	DH	N2	H3-JH
CR080362		AKGPSVPAARAEYFQH	G	PS	VPA	AR	AEYFQH
CR080363		AREGGLGYYYREWYFDL	E	GGL	GYYY	RE	WYFDL
CR080372		AKPDYGAEYFQH	-	P	DYG	-	AEYFQH
EK080902		AKEIVVPSAEYFQH	E	-	IVV	PS	AEYFQH

Table 54. Sequences of CDRH3 from Four HEL Binders in Numbering System of the Invention, According to the Numbering System of the Invention

Clones	[Tail]	[N1]			[DH]					[N2]			[H3-JH]										CDRH3 Length
	95	96	96A	96B	97	97A	97B	97C	97D	98	98A	98B	99E	99D	99C	99B	99A	99	100	101	102	
CR080362	G	P	S	-	V	P	A	-	-	A	R	-	-	-	-	A	E	Y	F	Q	H	14
CR080363	E	G	G	L	G	Y	Y	Y	-	R	E	-	-	-	-	-	W	Y	F	D	L	15
CR080372	-	P	-	-	D	Y	G	-	-	-	-	-	-	-	-	A	E	Y	F	Q	H	10
EK080902	E	-	-	-	I	V	V	-	-	P	S	-	-	-	-	A	E	Y	F	Q	H	12

The heavy chain chassis isolated were VH3-23.0 (for EK080902 and CR080363), VH3-23.6 (for CR080362), and VH3-23.4 (for CR080372). These variants are defined in Table 8 of Example 2. Each of the four heavy chain CDRH3 sequences matched a designed sequence from the exemplified library. The CDRL3 sequence of one of the clones (ED080902) was also determined, and is shown below, with the surrounding FRM regions underlined:

CDRL3: YYCQESFHIPYTFGGG.

In this case, the CDRL3 matched the design of a degenerate VK1-39 oligonucleotide sequence in row 49 of Table 33. The relevant portion of this table is reproduced below, with the amino acids occupying each position of the isolated CDRL3 bolded and underlined:

hassis	CDR Length	Junction type	Degenerate Oligonucleotide	SEQ ID	89	90	91	92	93	94	95	96	9
K1-39	9	1	CWGSAAWCATH CMVTABTCCTT WCACT		L**Q**	**E**Q	**S**T	**F**SY	**H**NPRST	**I**ST	**P**	F**Y**	**T**

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 and methods described herein. Such equivalents are intended to be encompassed by the scope of the following claims.

APPENDIX A

GI Numbers of Kappa Light Chains Used to Derive the VK Libraries

23868
32779
32810
33059
33144
33156
33170
33173
33183
33185
33189
33191
33195
33200
33202
33221
33227
33230
33233
33237
33268
33288
33290
33294
33296
33298
33300
33302
33304
33324
33330
33415
33416
33417
33418
33421
33422
33423
33424
33426
33647
33649
33655
33657
33659
33665
33669
33679
33683
33685
33756
2385488
2385490
2385492
2385494
2385495
2385497
2597932
2597935
2597937
2597943
2597946
2597948
2597950
2597952
2599531
2599533
2599535
2599545
2625059
2632152
2654047
2654051
2654055
2773084
2920359
2995674
2995676
2995678
2995680
2995682
2995688
2995690
3023134
3023136
3023138
3023140
3023142
3023144
3023146
3023148
3251385
3251387
3251389
3251391
3251744
3251749
3251983
3251985
3288824
3378165
3378177
16923194
16923202
16923208
17226623
17226631
17226635
17226639
17226643
17226645
17226655
17381491
17385013
17385015
17385017
17385019
17385021
17483729
18025561
18025563
18025573
18025575
18025577
18025579
18025581
18025583
18025585
18025587
18025589
18025591
18025593
18025595
18025597
18025599
18025603
18025605
18025607
18025611
18025613
18025617
18025621
18025623
18025627
18025629
18025635
18025639
18025641
18025645
18025651
18025653
18025655
18025657
58222611
58222613
58222615
58222617
58222619
58222621
58222623
58222625
58222627
58222629
58222631
58222633
58222635
58222637
58222639
58222641
58222643
58222645
58222647
58222649
58222651
58222653
58222655
58222657
58222659
58222661
58222663
58222665
58222667
58222669
58222671
58222673
58222675
58222677
58222679
58222681
58222683
58222685
58222687
58222689
58222691
58222693
58222695
58222697
58222699
58222701
58222703
58222705
58222707
58222709
58222711
70798854
70798856
70798858
70798860
70798862
70798866
70798868
70798872
70798874
70798878
70798880
70798882
70798884
70798886
70798888
70798890
70798892
70798894
70798896
70798898
70798900
70798902
70798904
70798906
70798914
70798916
70798918
70798920
70798922
70798926
70798928
70798930
70798934
70798936
70798940
70798942
70798946
70798948
70798950
70798952
70798954
70798956
71058688
71058704
71058712
71058717
71058719
71058721
71058723
71058725
71058727
98956311
98956323
98956325
98956327
98956337
98956341
98956343
98956349
98956355
98956357
98956365
98956375
98956379
98956381
98956383
98956400
98956404
98956406
98956414
98956418
98956422
98956426
98956428
98956430
98956432
98956436
98956440
99022977
99022979
99022981
99022983
99022985
99022987
99022989
99022991
99022993
99022995
99022997
99022999
99023002
99023004
99023006
99023008
99023010
99023012
99023474
99023476
99023478
99023480
99023482
99023484

34022
36657
37860
37909
38361
38362
38363
38367
38436
38438
38439
38440
38441
38442
38448
38485
38487
38489
38491
38493
38495
38497
38499
38501
38503
38505
178678
182338
182340
182342
182344
182346
182348
183962
183968
183972
185375
185377
185379
185381
185383
185385
185387
185389
185391
185393
185395
185397
185399
185401
185403
185415
185417
3378183
3451194
3603382
3603384
3603386
3603388
3603390
3603392
3603394
3603396
3641303
3641307
3644015
3644021
3746530
3747011
3747015
3821085
3821088
3901025
3928173
3928181
3928185
3928189
3928210
3928211
3928212
3928214
3928215
3928219
3928220
3928222
3928223
3928224
3928225
3928227
3928231
3928232
3928233
3928234
3928235
3928236
3928237
3928238
3928239
3928240
3928243
3928244
3928245
3928248
3928250
3928251
3928252
18025659
18025661
18025665
18025667
18025669
18025677
18025679
18025681
18025683
18025685
18025687
18025689
18025693
18025697
18025701
18025705
18025709
18025715
18025717
18025719
18092607
18092609
18092611
18092613
18092615
18092617
18092619
18092621
18092623
18307263
18307265
18307267
18307269
18307271
18307273
18307275
18307277
18307279
18307281
18307283
18307285
18307289
18307291
18307293
18626727
18626728
18626729
18626730
18632678
18698406
19170347
19701578
19744467
58222713
58222715
58222717
58222719
58222721
58222723
58222725
58222727
58222729
58222731
58222733
58222735
58222737
58222739
58222741
58222743
58222745
58222747
58222749
58222751
58222753
58222755
58222757
58222759
58222761
58222763
58222765
58222767
58222769
58222771
58222773
58222775
58222777
58222779
58222781
58222783
58222785
58222787
58222789
58222791
58222793
58222795
58222797
58222799
58222801
58222803
58222805
58222807
58222809
58222811
58222813
58222815
58222817
71058729
71058731
71482591
71482622
71482624
71482634
71482636
71482638
71482640
71482642
71482644
71482646
71482648
71482650
71482652
71482654
71792302
71792306
71792308
73532341
75707120
75707124
75707126
75707128
75707130
75707132
75707134
75707138
75707140
75707148
75707154
75707156
75707158
75707160
75707162
75707168
75707170
75707172
75707174
75707176
75707180
75707188
75707194
75707196
75707198
75707204
75707206
75707208
75707210
75707220
75707222
75707226
75707228
99025082
99025083
99025084
99025903
99025916
99026398
99026399
99026416
99026418
109240611
109240615
109240619
109240627
109240631
109240635
109240637
109240641
109240643
109240647
109240655
109240657
109240661
109240665
109240669
109240671
109240675
109240679
109240687
109240691
109240695
109240701
109240705
109240709
109240713
109240717
109240721
109240723
109240729
109240733
109240737
109240741
109240745
109240760
109240764
109240766
109240770
109241210
109241212
109241214
109241216
109241218
109241220
109241450

185419
185423
185427
185811
185813
185815
185816
185827
185829
185831
185833
185835
185837
185839
185841
185845
185847
185849
185855
185859
185862
185866
185868
185870
185872
185874
185880
185882
185884
185886
185888
185890
185892
185894
185896
185898
185904
185906
185908
185910
185912
185920
185922
185928
185934
185950
185980
185984
185987
185988
186008
186015
186017
3928253
3928254
3928257
3928258
3928259
3928260
3928261
3928263
3928264
3928265
3928266
3928267
3928276
3928277
3928278
3928279
3928280
3928283
3928287
3928288
3928289
3928290
3928291
3928293
3928294
3928295
3928296
3928297
3928298
3928299
3928301
3928302
3928303
3928304
3928308
3928309
3928310
3928312
3928315
3928316
3928317
3928318
3928319
3928320
3928321
3928323
3928324
3928325
3928326
3928327
3928329
3928330
3928331
19744471
19744475
19744479
19744487
19744491
19744495
19744499
19744503
19744507
19744511
19744515
19744519
19744523
19744527
19744531
19744535
19744539
19744543
19744547
19744551
19744555
19744559
19744563
19744567
19744571
19744575
19744579
19744583
19744587
20372497
20372499
20372501
20372503
20372505
20372507
20372509
20372511
20372513
20372515
20372517
20372519
20372521
20372523
20372525
20372527
20372529
20387057
20387059
20387061
21311286
21311288
21311294
21311296
58222819
58222821
58222823
58222825
58222828
58222830
58222832
58222834
58222836
58222838
58222840
58222843
58222845
58222847
58222849
58222851
58222853
58222855
58222857
58222859
58222861
58222863
58222865
58222867
58222869
58222871
58222873
58222875
58222877
58222879
58222881
58222883
58222885
58222887
58222889
58222891
58222893
58222895
58222897
58222899
58222901
58222903
58222905
58222907
58222909
58222911
58222913
58222915
58222917
58222919
58222923
58222925
58222927
75707230
75707232
75707234
75707236
75707238
75707240
75707242
75707244
75707246
75707248
75707250
75707262
75707264
75707268
75707270
75707272
75707274
75707276
75707278
75707282
75707284
75707292
75707298
75707300
75707302
75707304
75707306
75707316
75707318
75707322
75707324
75707334
75707338
75707340
75707362
75707368
75707370
75707372
75707374
75707378
75707382
75707384
75707386
75707398
75707406
75707408
75707410
75707412
75707416
75707418
75707420
75707422
75707424
109241549
109241551
109242373
109242377
109242379
109242381
109242383
109242385
109242387
109242389
109242395
109242399
109242401
109242403
109242409
109242411
109242417
109242419
109242421
109242423
109242425
109242427
109245190
109245192
109245194
109693080
109693082
109693084
109693094
109693096
109693100
109693102
109693110
109693112
109693114
109693116
109693118
109693120
109693135
109693137
109693139
109693144
109693146
109693148
109693150
109693152
109693154
109693157
109693159
109693165
109693167
109693169
109693171

186019
186040
186041
186042
186047
186199
186266
254719
257550
261239
265236
265240
298552
298560
298827
298829
299955
306919
306957
306959
306961
306963
306965
306967
306971
306980
306982
306984
306986
306988
306990
306992
306994
306996
306998
307000
348203
348205
348207
348211
386052
396631
397787
397789
397791
397793
397795
398490
398491
398492
404110
404112
404114
3928332
3928333
3928334
3928335
3928336
3928337
3928338
3928339
3928340
3928341
3928342
3928343
3928344
3928345
3928346
3928347
3928348
3928349
3928350
3928351
3928352
3928353
3928354
3928355
3928356
3928357
3928358
3928359
3928360
3928361
3928362
3928363
3928364
3928365
3928366
3928367
3928368
3928369
3928370
3928371
3928372
3928373
3928374
3928375
3928376
3928377
3928378
3928379
3928380
3928381
3928382
3928383
3928384
21311318
21311322
21669062
21669064
21669066
21669068
21669070
21669072
21669074
21669076
21669078
21669080
21669082
21669084
21669086
21669088
21669090
21669092
21669094
21669096
21669098
21669100
21669102
21669104
21669106
21669108
21669110
21669112
21669114
21669116
21669118
21669120
21669122
21669124
21669126
21669128
21669130
21669132
21669134
21669136
21669138
21669140
21669142
21669144
21669146
21669148
21727250
21998806
21998808
21998810
21998812
21998814
21998816
58222929
58222931
58222933
58222935
58222937
58222939
58222941
58222943
58222945
58222947
58222949
58222951
58222953
58222955
58222957
58222959
58222961
58222963
58222965
58222967
58222969
58222971
58222973
58222975
58222977
58222979
58222981
58222983
58222985
58222987
58222989
58222991
58222993
58222995
58222997
58222999
58223001
58223003
58223005
58223007
58223009
58223011
58223013
58223015
58223017
58223019
58223021
58223023
58223025
58223027
58223029
58223031
58223033
75707426
75707428
75707430
75707432
75707434
75707444
75707446
75707448
75707454
75707460
75707462
75707464
75707472
75707476
75707500
75707502
75707504
75707506
75707508
75707510
75707514
75707516
75707518
75707520
75707522
75707524
75707526
75707528
75707530
75707534
75707536
75707540
75707542
75707544
75707546
75707548
75707550
75707552
75707586
75707598
75707600
75707602
75707604
75707618
76058957
76252624
76252626
76252630
76252632
76252634
76252636
76252638
76252640
109693177
109693179
109693181
109693183
109693187
109693189
109693201
109693203
109693206
109693210
109693216
109693218
109693220
109693222
109693228
109693230
109693232
109693235
109693237
109693239
109693241
109693249
109693253
109693255
109693261
109693264
109942421
109942431
110290934
110610132
110624509
110657101
110657103
110657105
110657107
110657109
110657111
110657113
110657115
110657123
110657124
110657125
110657158
110657159
110657160
110657161
110657162
110657163
110657164
110657165
110657166
110657167
110657168

408365
409042
414035
415651
415710
415955
415957
415959
415961
415963
415965
415967
415969
415971
416329
416331
416333
416335
416337
430845
431039
431040
431041
431042
431043
431044
431045
431046
431047
431048
431049
431051
431052
431053
431067
431069
431071
431073
431075
431077
431079
431081
431083
431085
431087
431089
433889
436562
440153
441312
441314
441316
441318
3928385
3928386
3928387
3928388
3928389
3928390
3928391
3928392
3928393
3928394
3928395
3928396
3928397
3928398
3928399
3928400
3928401
3928402
3928403
3928404
3928405
3928406
3928407
3928408
3928409
3928410
3928411
3928412
3928413
3928414
3928415
3928416
3928417
3928418
3928419
3928420
3928421
3928422
3928423
3928424
3928425
3928426
3928427
3928428
3928430
3928431
3928432
3928433
3928434
3928435
3928436
3928437
3928438
21998818
21998820
21998822
21998824
21998826
21998830
21998832
22086572
22086575
22086581
22086587
22086593
22091617
22214019
22214023
22297542
22556681
22556683
22556684
22607990
22620896
22620899
22640510
22640512
22640513
22642789
22642790
22642791
22642808
22642809
22642810
22642811
22643188
22643190
22643192
22643196
22647625
22647633
23194480
23194500
23225992
23225994
23225996
23234613
23320663
23342423
23343554
24412754
24412756
24412758
24474081
24850297
26985941
58223035
58223037
58223039
58223041
58223043
58223045
58223047
58223049
58223051
58223053
58223055
58223057
58223059
58223061
58223063
58223065
58223067
58223069
58223071
58223073
58223075
58223077
58223079
58223081
58223083
58223085
58223087
58223089
58223091
58223093
58223095
58223097
58223099
58223101
58223103
58223105
58223107
58223109
58223111
58223113
58223115
58223117
58223119
58223121
58223123
58223125
58223127
58223129
58223131
58223133
58223135
58223137
58223139
76252642
76252644
76252646
76781673
77378090
77378092
77378094
77378096
77378098
77378100
77378102
77378105
77378107
77378109
77378111
77378135
77378137
77378139
77378141
77378143
77378145
77378147
77378149
77378151
77378153
77378155
77378157
77378159
77378161
77378163
77378165
77378167
77378169
77378172
77378174
77378176
77378224
77378225
77378228
77378230
77378234
77378236
77378237
77378239
77378241
77378245
77378247
77378249
77378251
77378253
77378255
77379405
77379407
110657169
110657170
110657171
110657172
110657173
110657174
110657175
110657176
110657177
110657178
110657179
110657180
110657181
110657182
110657183
110657184
110657185
110657186
110657187
110657188
110657189
110657230
110657232
110657234
110657236
110657238
110657240
110657242
110657244
110657246
110657248
110657250
110657252
110657254
110657256
110657258
110657615
110657617
110657619
110657621
110657624
110657676
110657678
110657728
110657730
110658341
110660158
110660166
110660174
112184495
112184497
112184499
112184501

441320
441322
441324
441330
441332
441334
441336
441338
441342
441344
441346
441348
441350
441352
441354
441356
441358
441360
441364
441366
441368
441370
441372
441374
441376
441378
441380
441382
441384
441386
441388
441390
441392
441394
441396
441398
441400
441402
441408
441412
441414
441416
441418
441422
441424
441426
441428
441430
441432
441434
441436
441440
441444
3928439
3928440
3928441
3928442
3928443
3928444
4100379
4100381
4100383
4103644
4103662
4103664
4103666
4103674
4128063
4139195
4139197
4139199
4139201
4323178
4323182
4323186
4323194
4323809
4323811
4323813
4323821
4323823
4323825
4323829
4323831
4323833
4323839
4323841
4323845
4323847
4323849
4323851
4323853
4323855
4323857
4323859
4323861
4323863
4323865
4323869
4323871
4323873
4323875
4323877
4323881
4323883
4323885
27368974
27368976
27368978
27368981
27368983
27368986
27368991
27368993
27368997
27368999
27369001
27369003
27369007
27369009
27369011
27818830
27867541
27873542
27875080
27875088
27875191
27875199
28611056
28848873
28883544
28883548
28883550
29650328
29650334
29650337
29650339
29725711
29725713
29725715
29725717
29725719
29725721
29725723
29725725
29725727
29725729
29725731
29725733
30026987
30258344
30258346
30793253
30793255
30793257
30793259
30793261
30793263
30793265
58223141
58223149
58223151
58223153
58223155
58223157
58223159
58223161
58223163
58223165
58223167
58223169
58223171
58223173
58223175
58223177
58223179
58223181
58223183
58223185
58223187
58223189
58223191
58223193
58223195
58223197
58223199
58223201
58223203
58223205
58223207
58223209
58223211
58223213
58223215
58223217
58223219
58223221
58223223
58223225
58223227
58223229
58223231
58223233
58223235
58223237
58223239
58223241
58223243
58223245
58223247
58223249
58223251
77379409
77379412
77379414
77379416
77379418
77379420
77379422
77379425
77379427
77379429
77379431
77379433
77379435
77379437
77379439
77379441
77379443
77379445
77379447
77379449
77379457
77379459
77379461
77379463
77379477
77379479
77379481
77379483
77379485
77379487
77379489
77379491
77379493
77379495
77379497
77379499
77379501
77379503
77379505
77379507
77379509
77379511
77379513
77379515
77379517
77379519
77379521
77379523
77379525
77379527
77379529
77379545
77994607
112184503
112184505
112184507
112184509
112184511
112184513
112189154
112191695
112191699
112703827
112708249
112708250
112711584
112712351
112712352
112712353
112712354
112712355
112712356
112712357
112712358
112712359
112712360
112712361
112712362
112712363
112712364
112712365
112712366
112712367
112712368
112712369
112712370
112712371
112712372
112712373
112712374
112712375
112712376
112712377
112712378
112712379
112712380
112712381
112712382
112712383
112727205
112727206
112727207
112727208
112727209
112727210
112727211

441446
441448
452060
452061
452062
452063
459655
460858
472970
472971
472972
472973
472974
472975
472976
487826
487827
493148
493149
493150
496044
496046
496048
496050
496053
496055
496059
496061
496063
496065
496071
496073
506420
506424
510839
510841
510843
510845
514428
514429
514430
514431
514432
514433
514434
515780
516137
516187
516198
516213
516249
516265
516316
4323887
4323889
4323891
4323893
4323895
4323897
4323899
4323901
4323903
4323905
4323907
4323909
4323911
4323913
4323915
4323923
4323927
4323929
4323931
4323933
4323935
4323937
4323939
4323941
4323945
4323947
4323949
4323951
4323953
4323955
4323957
4323959
4323961
4323963
4323965
4323983
4323989
4323993
4323997
4323999
4324005
4324007
4324009
4324011
4324013
4324019
4378181
4378183
4378185
4378187
4378189
4378191
4378193
30793565
30793567
30793569
30793571
30793573
30841928
30841931
30841933
30841935
30841939
30841943
30841945
30841947
31879463
31879464
31879467
31879468
31879471
31879472
33021483
33044572
33044573
33044574
33044582
33044586
33051527
33051528
33070272
33070283
33070284
33083474
33083476
33083477
33083478
33083479
33083480
33083481
33083482
33083483
33085842
33235609
33235611
33235613
33235615
33235617
33235619
33235621
33235623
33235625
33235627
33235629
33235631
33235633
58223253
58223255
58223257
58223259
58223261
58223263
58223265
58223267
58223269
58223271
58223273
58223275
58223277
58223279
58223281
58223283
58223285
58223287
58223289
58223291
58223293
58223295
58223297
58223299
58223301
58223303
58223305
58223307
58223309
58223311
58223313
58223315
58223317
58223319
58223321
58223323
58223325
58223327
58760238
59890568
59890571
59894819
60392126
60616327
60616352
60650119
60650123
60734312
61697118
61853816
61970154
61970158
61970160
77994611
77994615
77994619
78629976
78629977
78629978
80750467
80975580
80975600
80975604
80975616
80975618
80975638
80975642
80975644
81020146
81020229
81020258
81239122
81251581
81251585
82794837
83410334
83697271
83959521
83959523
83959525
83959937
83959939
83964685
83964762
83964764
83964766
83964768
83966574
83966576
83966578
83966655
83966657
83966659
83966661
83966663
83966665
83966667
83970756
83970763
83970769
83970772
84659318
84659320
84660715
84660717
84660719
112727212
112727213
112727214
112727215
112727216
112727217
112727218
114155738
114155883
114155884
114156208
114207907
114385493
114385505
114385507
114385509
114385511
114385513
114385515
114385517
114385521
114385537
114385539
114385541
114385543
114385545
114385547
114385549
114385551
114385553
114385567
114385569
114385571
114385573
114385575
114385579
114385581
114385583
114385585
114385587
114385589
114385591
114385593
114385595
114385597
114385599
114385601
114385603
114385605
114385607
114385609
114385611
114385613

545722
557650
557651
560677
560678
560841
560843
575228
575236
575240
575257
575261
587143
587245
587323
587325
587327
587329
587331
587333
587335
587337
587341
587343
587345
587347
587349
587351
587353
598165
598167
598170
598172
601979
601982
601984
609002
609004
619259
623043
624874
632983
632985
632987
633227
642581
681896
681899
685029
693862
722413
722417
722419
4378195
4378197
4378199
4378201
4378203
4378207
4378209
4378211
4378213
4378215
4378217
4378221
4378223
4378225
4378227
4378229
4378233
4378237
4378239
4378243
4378245
4378247
4378249
4378251
4378253
4378255
4378259
4378261
4378265
4378267
4378269
4378271
4378273
4378275
4378279
4378281
4378283
4378287
4378291
4378293
4378295
4378297
4378299
4378301
4378303
4378305
4378307
4378309
4378313
4378315
4378317
4378319
4378323
33304656
33304658
33304661
33304663
33355480
33868634
33868636
33868638
33868640
33868642
33868644
33868646
37287525
37605051
37694620
37694622
37694624
37694626
37694628
37694630
37694632
37694634
37694636
37694638
37694640
37694642
37694644
37694646
37694648
37694650
37694654
37694660
37694662
37694664
37694666
37694668
37694670
37694672
37694674
37694676
37694678
37694680
37694682
37694684
37694686
37694688
37694690
37694692
37694694
37694696
37694698
37694700
37694702
61970164
61970168
61970172
61970176
61970180
61970184
61970192
61970194
61970198
61970202
61970206
61970228
62001845
62120916
62120917
62120918
62120919
62120920
62120921
62120922
62120923
62120924
62120925
62120926
62120927
62120929
62120931
62120932
62120933
62120934
62120935
62120938
62120939
62120940
62120941
62120943
62120944
62120945
62120946
62120947
62120948
62120949
62120950
62120951
62120952
62120953
62120954
62120955
62120956
62120957
62120958
62120959
62120960
84660720
84660721
84660722
84660723
84660725
84797793
84797795
84797797
84797799
84797801
84797803
84797805
84797807
84797823
84797825
84797827
84797857
84797861
84797883
84797915
84797929
84797959
84797961
84797963
84797979
84797981
84797985
84798001
84798003
84798005
84798007
84798009
84798011
84798033
84798035
84798055
84798057
84798059
84798061
84798063
84798103
84798107
84798115
84798117
84798147
84798149
84798167
84798169
84798171
84798173
84798175
84798177
84798179
114385615
114385617
114385619
114385621
114385623
114385625
114385627
114385629
114385631
114385633
114385635
114385645
114385647
114385649
114385651
114385653
114385655
114385659
114385661
114385663
114385665
114385669
114385671
114385673
114385675
114385677
114385679
114385681
114385683
114385685
114385687
114385689
114385691
114385693
114385699
114385701
114385703
114385705
114385707
114385709
114385711
114385713
114385715
114385717
114385719
114385721
114385723
114385725
114385727
114385729
114385731
114385744
114385746

722421
722423
722425
722427
722429
722431
722433
722435
722437
722439
722441
722443
722455
722461
722463
722465
722467
722469
722471
722473
722475
722477
722479
722483
722485
722487
722489
722493
722495
722497
722503
722505
722511
722513
722515
722521
722523
722525
722529
722531
722535
722537
722539
722541
722543
722545
722549
722553
722555
722557
722559
722561
722569
4378325
4378327
4378331
4378333
4378335
4378337
4378339
4378341
4378343
4378345
4378347
4378349
4378351
4378353
4378359
4378361
4378363
4378365
4378367
4378369
4378371
4378373
4378375
4378377
4378379
4378383
4378385
4378387
4378389
4378391
4378393
4378395
4378397
4378399
4378401
4558868
4680172
4759539
4759543
4759547
4759551
4759555
4759563
4759567
4759575
4759579
4759583
4759587
4759591
4759595
4759599
4761194
4761281
37694704
37694706
37694708
37694710
37702652
37732215
37780362
39103877
39103879
39103881
39103883
39103885
39103887
40231616
40288410
40288412
40288414
40288416
40288418
40388582
40388585
40388592
40388599
40647131
40784425
40784429
40795876
42541061
42541069
42794782
42794786
44829186
45111420
45386482
46016047
46093898
46093902
46093906
46093910
46575858
47078185
47154907
47154909
47154911
47154913
47154915
47154917
47154919
47154921
47271269
47271271
47271273
47271275
62120961
62120962
62199500
62421462
62421466
62720427
62720431
62720436
62720442
62720444
62720446
62720452
62720454
62720473
62720475
62720477
62720483
62860940
62860955
62860957
62860959
62860961
62860963
62860965
62860981
62860983
62860987
62860989
62860991
62860994
62860996
62861000
62861002
62861004
62861012
62861015
62861017
62861019
62861022
62861024
62861029
62861031
62861037
62861041
62861045
62861054
62868475
62868477
62868479
62999493
63102866
63102872
63102874
84798181
84798183
84798197
84798199
84798201
84798203
84798213
84798215
84798217
84798219
84798241
84798249
84798255
84798257
84798267
84798269
84798271
84798273
84798275
84798277
84798279
84798295
84798309
84798321
84798323
84798325
84798327
84798343
84798345
84798347
84798349
84798351
84798364
84798366
84798370
84798372
84798374
84798377
84798381
84798383
84798386
84798388
84798390
84798397
84798407
85632219
85642735
85644222
85644224
85644226
85644228
85644230
85644232
114385748
114385750
114385752
114385756
114385774
114385776
114385778
114385780
114385782
114385804
114385806
114385808
114385921
115268711
115268713
115268880
115268892
115268894
115268896
115268898
115268900
115268902
115268904
115268906
115270875
115270877
116543556
116543560
116543564
116546686
116546688
116551153
116551156
116551162
116551171
116551175
116551179
116551183
116551188
116551192
116551201
116551207
116551216
116551226
116551231
116551235
116551239
116551244
116551249
116551258
116551313
116551317
116551321

722571
722573
722581
722585
722587
722591
722593
722599
722601
722603
722605
722607
722609
722615
732737
732739
732741
732743
732745
732747
758588
758598
758600
762823
773589
790442
790450
790794
790802
790810
791015
791019
791023
791027
791031
791035
809552
809553
809554
845515
845517
845519
845521
845523
845525
845527
845529
845531
845533
845535
854111
871275
871819
4761283
4837686
4837688
4837690
4837692
4837694
4837696
4837698
5006350
5006354
5006356
5006358
5006360
5019510
5019512
5019514
5019522
5019524
5019526
5019538
5081714
5081716
5081718
5081720
5081722
5102680
5419682
5419684
5419700
5419702
5419704
5419706
5419708
5419710
5419712
5419731
5419738
5419740
5524134
5524140
5524142
5524144
5524146
5524148
5524150
5566507
5578779
5578781
5578783
5578785
5578787
5578789
5578791
47271277
47271279
47271281
47271283
47271285
47271287
47271289
47271291
47271295
47271297
47271299
47271307
47271309
47271311
47271313
47271315
47271317
49073024
49073036
50199324
50199334
50831237
50844518
50844522
50844526
50844536
50844540
50844548
50844552
50871685
50871687
50898144
50898148
50898150
50898152
50898154
50898158
50898160
50898162
50898164
50898170
51103388
51103390
51103392
51103394
51103396
51103398
51103400
51103402
51103404
51103406
51103408
51103410
63102876
63102880
63102882
63102888
63102892
63102898
63102900
63102902
63102904
63102906
63102908
63102910
63102912
63102916
63102920
63102922
63102924
63102928
63102938
63102940
63102942
63102954
63102962
63102964
63102966
63102968
63102970
63102972
63102974
63102976
63102980
63102986
63102988
63102992
63102994
63102996
63102998
63103012
63103014
63103030
63103032
63103034
63103040
63103044
63103046
63103048
63103054
63103056
63103070
63103072
63103076
63103078
63103086
85644600
85644602
85644604
85650161
85650163
85650165
85650167
85650169
85650171
85650173
85650175
85650177
85650179
85650276
85650278
85650280
85657010
85658337
85658632
85660488
85660492
85660494
85660497
85660498
85660502
86439043
86439047
86439051
86439053
86439057
86439061
86439063
86439071
86439075
86439081
86439147
86439151
86439153
87298995
87298999
87299001
87299003
87299007
87299009
87299011
87299015
88496317
88496922
90092372
90092373
90092374
90092387
90092910
116551325
116551329
116551333
116551337
116551341
116551347
116551351
116551369
116551373
116551377
116551381
116551404
116551413
116551418
116551422
116551427
116551431
116551436
116551446
116551452
116551772
116551776
116551780
116551785
116551790
116553242
116555276
116555819
116555821
116555823
116559889
116560960
116634471
116634475
116795086
117576090
118143176
118143178
118147088
118147090
118147092
118147094
118147096
118147098
118147100
118147102
118147104
118147106
118147108
118147110
118147112
118147114
118147116

871823
882261
882263
882265
882267
882269
882271
882273
882275
882277
882279
882281
882283
882285
882287
882289
882291
882293
882295
882297
882299
882301
882303
882305
882307
882309
882311
882313
882315
882317
882319
882321
882323
882325
882327
882329
882331
894090
904629
913352
929640
929642
944925
950049
973411
973415
999107
1020008
1020012
1020016
1070309
1070313
1070315
5578793
5578795
5578797
5578799
5578801
5578803
5578805
5578807
5578809
5578811
5578815
5690395
5690399
5690403
5709454
5731228
5731232
5731236
5731242
5731252
5921608
5921610
5921614
5921618
5921620
5921622
5921624
5921626
5921640
6110569
6179861
6179863
6179865
6179867
6179869
6492198
6492200
6492202
6492204
6648587
6649889
6649895
6708204
7012704
7012706
7024356
7160978
7673384
7673388
7673392
7745134
8250280
8777870
51103412
51103414
51103416
51103418
51103420
51103422
51103424
51103522
51103526
51103528
51103532
51103534
51103536
51103538
51103540
51103542
51103544
51103546
51103548
51103550
51103552
51103554
51103556
51103558
51103560
51103562
51103564
51103566
51103568
51103570
51851021
51949938
53988135
53988137
54034484
54145422
54145426
54145440
54781098
54781100
54781102
54781104
54781106
54781108
54781110
54781112
54781126
54781129
54781202
54781204
54781206
54781208
54781213
63103096
63103098
63103106
63103108
63103110
63103112
63103114
63103116
63103118
63103120
63103140
63103142
63103144
63103146
63103148
63103150
63103154
63103156
66096574
66096603
66096637
66711101
66711102
66711103
66711104
66711105
66711106
66711107
66711108
66711109
66711110
66711111
66711112
66711114
66711116
66711117
66711118
66711119
66711120
66711123
66711124
66711125
66711126
66711128
66711129
66711130
66711131
66711132
66711133
66711134
66711135
66711136
66711137
90092911
90092912
90092913
90823178
90823182
90823186
90823190
90823196
90823198
90994745
90994747
90994751
92115496
92115497
92130102
92130103
92131782
92131783
92131784
92131785
92133663
92133665
92137567
92140334
92140336
92141530
92155949
92157443
92157445
92157453
92157459
92157461
92158828
92158980
92161545
92249233
92298212
92298539
92315622
92315624
92315626
92315628
92332837
92332841
92348102
92348670
92349881
92360819
92370888
92381676
92496960
92520581
92520583
118147118
118147120
118147122
118147125
118147127
118425771
118425773
118425775
118490144
118490148
118490152
118490156
119359417
119836694
119836767
119838997
119839065
119839355
119839523
119841342
119841388
119841425
119841512
121309186
124042790
124042792
124042815
126146964
126146965
126146966
126147776
126147812
126147817
126147952
126147954
126147956
126152193
126152196
126633956
126633957
126633958
134125852
134125853
134125854
134128019
134269772
134273023
145850477
145850518
145850519
145850520
145850521
145850522

1070317
1070321
1070325
1070327
1070347
1136554
1136556
1208913
1235764
1235766
1235768
1235770
1235772
1235774
1245380
1245382
1255605
1255607
1255608
1255609
1255612
1292860
1292862
1353813
1353815
1353817
1353819
1353821
1353825
1353827
1353831
1370131
1370135
1370137
1495627
1495628
1495629
1495630
1495631
1495632
1495633
1495634
1495635
1495637
1495638
1495639
1495640
1495641
1495642
1495643
1495644
1495645
1495646
8777874
8777878
8777880
8777884
8777888
8777890
8777892
9295278
9295280
9295282
9295284
9295286
9295290
9295292
9295296
9295298
9295300
9437312
9927567
9928208
9968441
9968443
9968486
9968488
9968490
9968492
9968494
9968496
9968498
9968500
9997457
10636524
11229436
11343336
11343337
11876718
11876734
11876735
11876736
11876737
11876738
11876739
11876740
11876741
11878173
11878175
11878177
11992075
11992193
12003249
12003251
12003253
12003255
54781216
54781218
54781220
54781223
54781225
54781227
54781229
54781231
55274149
55274153
55274159
55274163
55824376
56118076
56118080
56292538
56294837
56294841
56399565
56609227
56609228
56609229
56609230
56609232
56609235
56742105
56742106
58003567
58003568
58003569
58003570
58003571
58003572
58003573
58003587
58003588
58003589
58003608
58003609
58003610
58003611
58003612
58003613
58003614
58003615
58003616
58003618
58003619
58003620
58003622
58003623
58003624
58003625
66711138
66711139
66711140
66711141
66711142
66711143
66711144
66711145
66711146
66711147
67509857
67509861
68148126
68148140
68148142
68148144
68148150
68148152
68148154
68148158
68148160
68148164
68148166
68148174
70797818
70797820
70797822
70797824
70797826
70797828
70797830
70797832
70797834
70797836
70797838
70797842
70797844
70797846
70797850
70797852
70797854
70797856
70797858
70797860
70797866
70797870
70797872
70797874
70797876
70797878
70797884
70797886
70797888
92520584
92520586
92575636
92589636
92589637
92589638
92589639
92589640
92589641
92589642
92589643
92589644
92589645
92589646
92589647
92589648
92589649
92589650
92589651
92589652
92589653
92589656
92600475
92600479
92600487
92607622
92667306
92667307
92667308
92667309
92667310
92667329
92667331
92798195
92798196
92798197
92798198
92798199
92798218
92798220
92824835
92834676
92835832
92835834
92835836
92839400
92839402
92839403
92839404
92839405
92839406
92839407
92839408
145850523
145850524
145850525
145850526
145850527
145850528
145850529
145850530
145850531
145850532
145850533
145850534
145850535
145850536
145850537
145850558
145850561
145850563
145854440
145856824
145859735
148355517
148355518
148355519
148355520
148355521
148355522
148355523
148355524
148355525
148355526
148355527
148355528
148355529
148355530
148355531
148355532
148355533
148540957
148578450
148578452
148578454
148578455
148578456
148578457
148578458
148578460
149849068
149849080
149849084
149849088
150447881
150447883

1495647
1495648
1495649
1495650
1495651
1495652
1532001
1532002
1532027
1552277
1552283
1552285
1552287
1552291
1552295
1552299
1552319
1561601
1561605
1561607
1561609
1561611
1572702
1572704
1572706
1572708
1572710
1657324
1657326
1657328
1673592
1673602
1710418
1770403
1770415
1773056
1778125
1785869
1785873
1785877
1800286
1813653
1813655
1813657
1834498
1834563
1834564
1835872
1835873
1839291
1864110
1864112
1864114
12003257
12655491
12655493
12655500
12655502
12655504
12655519
12655521
12655525
12655527
12655529
12655531
12655541
12655558
12655565
12655567
12655569
12655643
12655655
12655662
12655665
12655672
12655713
12655723
12655730
12655732
12655736
12655738
12655740
12655748
12655751
12710669
12710671
12734084
12734089
12750933
12836990
12957385
12957387
13170940
13170944
13170948
13171333
13171339
13171341
13171343
13447996
13448000
13448002
13448004
13448006
13448010
13448012
58003626
58003627
58003628
58003629
58003630
58003631
58003632
58003633
58003634
58003656
58003657
58003658
58003659
58003660
58003661
58032596
58032603
58032606
58194104
58194120
58194136
58202701
58202709
58202711
58202713
58202715
58202717
58202719
58202721
58202723
58202725
58202727
58202729
58202733
58202735
58202737
58222454
58222456
58222458
58222460
58222462
58222464
58222466
58222468
58222470
58222473
58222476
58222478
58222480
58222482
58222484
58222487
58222489
70797890
70797894
70797898
70798601
70798603
70798605
70798607
70798609
70798611
70798613
70798615
70798617
70798619
70798621
70798623
70798627
70798629
70798631
70798633
70798635
70798637
70798639
70798641
70798643
70798645
70798649
70798653
70798655
70798657
70798659
70798661
70798667
70798669
70798671
70798673
70798675
70798677
70798679
70798681
70798683
70798685
70798687
70798690
70798692
70798694
70798696
70798698
70798700
70798702
70798706
70798708
70798710
70798712
92839409
92845038
92845490
92845651
92855396
92855400
92855404
92855408
92855412
92855416
92855420
92855424
92855428
92855432
92855436
92855441
92855444
92856854
92856855
92856859
92857001
92857003
92857012
92857016
92857018
92858156
92861312
92861313
92861314
92862784
92875826
92878541
92878543
92878545
92903931
92905358
92905360
92905362
94034254
94034257
94034261
94034264
94034267
94034271
94034285
94034316
94034339
94034342
94034384
94034387
94034390
94034393
94035272
150447885
150447887
150450134
150450135
150450136
150450137
150450138
150450139
150450140
150450636
150453145
150453147
150453149
150453151
150453153
150453154
150453155
150453156
150453157
150453159
150453161
150453163
150453165
150453167
150453169
150453171
150453174
150453213
150453216
153590356
153590359
153590361
153590363
153590365
153590367
153590371
156149223
156149224
156149225
156229617
156557387
156557389
156557391
156557393
156557399
156557403
156557405
156557407
156557411
156562058
157087534
157896695
157896697

1864116
1864118
1864136
1864138
1890131
1890133
1905798
1905937
1905941
1911732
1922370
1922438
1922466
1922501
1922528
1922535
1922602
1922618
1922645
1922679
1922796
1922805
1932772
1943727
2058533
2058535
2058678
2072271
2072273
2072279
2072981
2078359
2078371
2078373
2169989
2169990
2172285
2173403
2175768
2175852
2175867
2218123
2239113
2239115
2253439
2266632
2266634
2291087
2293965
2293967
2306827
2306829
2345025
13448016
13448018
13448022
13549147
13785652
13939245
13939277
13939331
13991697
14150696
14150698
14290262
14573212
14573214
14573216
14573218
14573220
14573222
14573226
14573254
14573256
14573258
14573260
14573262
14573264
14573266
14573268
14573270
14573272
14573274
14573276
14573278
14588864
14588866
14588868
14588870
14588872
14597098
14597112
14597124
14597127
14625743
14625918
14626493
14716957
14716961
14716969
14716971
14716973
15011457
15099974
15277619
15419020
58222491
58222493
58222497
58222499
58222501
58222503
58222505
58222507
58222509
58222511
58222513
58222515
58222517
58222519
58222521
58222523
58222525
58222527
58222529
58222531
58222533
58222535
58222537
58222539
58222541
58222543
58222545
58222547
58222549
58222551
58222553
58222556
58222558
58222560
58222562
58222564
58222566
58222568
58222570
58222572
58222575
58222577
58222579
58222581
58222583
58222585
58222587
58222589
58222591
58222593
58222595
58222597
58222599
70798716
70798718
70798720
70798722
70798724
70798732
70798734
70798736
70798738
70798742
70798744
70798750
70798752
70798758
70798760
70798764
70798766
70798768
70798770
70798772
70798774
70798776
70798778
70798780
70798782
70798784
70798786
70798788
70798792
70798794
70798796
70798798
70798800
70798802
70798804
70798806
70798808
70798810
70798812
70798814
70798816
70798818
70798820
70798824
70798826
70798828
70798830
70798832
70798834
70798836
70798838
70798840
70798842
94035284
94035289
94035298
94035300
94035312
94469910
94469912
94469914
94469922
94469924
94469926
95007504
95007510
95007512
95007514
95007516
95007518
95007520
95007522
95007524
95007526
95007528
95007530
95007532
95007534
95007536
95007538
95007540
95007542
95007544
95101759
95101761
95101767
95101769
95101777
98956195
98956209
98956219
98956223
98956232
98956244
98956249
98956255
98956261
98956263
98956271
98956277
98956279
98956281
98956285
98956289
98956291
98956293
157903220
158055245
158055254
158055268
158055282
158055285
158055288
158058441
158731523
158731524
158731525
158731526
158731527
158731528
158731529
158731530
158731531
158731532
158731533
158731534
158731536
158731538
158731539
158731540
158731541
158731542
158731545
158731546
158731547
158731548
158731550
158731551
158731552
158731553
158731554
158731555
158731556
158731557
158731558
158731559
158731560
158731561
158731562
158731563
158731564
158731565
158731566
158731567
158731568
158731569
158744132
158744140
158744148

2345029
2345031
2345033
2385484
2385486
15859220
15986229
16508167
16554974
16923186
58222601
58222603
58222605
58222607
58222609
70798844
70798846
70798848
70798850
70798852
98956299
98956301
98956303
98956305
98956307
158744156
158744164
158746355
158746363
158746371

APPENDIX B

GI Numbers of Lambda Light Chains Used to Derive the Vλ Libraries

31454
32808
32812
33335
33368
33383
33387
33412
33429
33431
33433
33703
33711
37918
37920
37922
37923
38359
38360
38364
38365
38366
38368
186078
186080
186082
186084
186086
186088
186090
186092
186094
186096
186097
186111
186162
186164
186168
186170
186172
186175
298556
405223
405227
409040
409041
409043
433485
434041
434045
439514
3142529
3142531
3142533
3142535
3142537
3142539
3142541
3142543
3142545
3142547
3142549
3142553
3142556
3142558
3142562
3142564
3142566
3142569
3142573
3142577
3142579
3142581
3142583
3142585
3142587
3142589
3142591
3142593
3142595
3142597
3142599
3142601
3142603
3142612
3142614
3142616
3142618
3142620
3142649
3142651
3142653
3142656
3142658
3142660
3142662
3142668
3142670
3142672
3142674
3142676
3142678
4566076
4566078
4566082
4566084
4566086
4566088
4566090
4566092
4566094
4566096
4566098
4566101
4566105
4732059
4761253
4761255
4761257
4761259
4761261
4761263
4761265
4761267
4761269
4761271
4761273
4761277
4761279
4927957
5019504
5019506
5019516
5019518
5019520
5019528
5019530
5019532
5019534
5019536
5174362
5174364
5174366
5174378
5524086
5524106
5524108
5524118
5524122
5524132
5578817
5578819
5578823
9968397
9968401
9968403
9968405
9968409
9968411
9968413
9968415
9968417
9968419
9968421
9968423
9968425
9968427
9968429
9968433
9968435
9968437
9968439
10636511
10636514
10636518
10636521
10636527
11992185
11992187
11992189
11992191
11992195
11992197
11992199
11992201
12666922
12666924
12666926
12666928
12666930
12666932
12666934
12666936
12666938
12666940
12666942
12666944
12666946
12666948
12666952
12666954
12666956
12666958
12666960
51103608
51103612
51103614
51103616
51490956
54781261
61815560
62720404
62720406
62720408
62720412
62860947
62860950
62860967
62860969
62860971
62860973
62860975
62860977
62860979
62860985
62861006
62861008
62861010
62861047
62999489
62999497
62999501
62999509
70888031
70888035
70888037
70888041
70888043
70888045
70888047
70888049
70888051
70888053
70888055
70888057
70888059
70888061
70888063
70888065
70888067
70888069
70888071
70888073
70888075
70888077
77379760
77379824
77379826
77379828
77379830
77379832
77379834
77379836
77379838
77379840
77379842
77379846
77379848
77379850
77379855
77379857
77379859
77379861
77379863
77379865
77379867
77379869
77379871
77379875
77379877
77379879
77379882
77379884
77379886
77379888
77379890
77379894
77379896
77379900
77379908
77379910
77379912
77379916
77379918
80975584
80975588
80975598
80975622
80975628
80975632
80975636
81020028
81020064
86438995
86439001
86439005

439516
441251
460854
460856
460860
465157
465167
465171
465175
469249
483911
487824
487825
487828
493153
506426
506428
515765
532599
532600
532603
560845
575230
575238
575242
685021
773591
871362
987068
987076
998390
998394
1055278
1070329
1070341
1070349
1143195
1200068
1235776
1235778
1235780
1235782
1255606
1255610
1255611
1255613
1552313
1561599
1770407
1864134
1864140
1864142
1864144
3142680
3142684
3153359
3153361
3153365
3153366
3153368
3153374
3153376
3335577
3335579
3335585
3335587
3335591
3388046
3388048
3388050
3388054
3388056
3388058
3388060
3388062
3388064
3388066
3388070
3388072
3388074
3388080
3747019
3821077
3821078
3821079
3821080
3821081
3821082
3821083
3821084
3821086
3821087
3821089
3821090
3821091
3821092
3821093
3821094
3821095
3821096
3821097
4103646
4103648
4103650
4103652
4103654
5578825
5578827
5578829
5578831
5578833
5911837
6492194
6492196
6492206
6492208
6492210
6492212
6643078
6643082
6643086
6643088
6643090
6643098
6643104
6643106
6643114
6643118
6643120
6643124
6643126
6643128
6643136
6643138
6643154
6643156
6643158
6643162
6643168
6643170
6643172
6643176
6643178
6643180
6643182
6643184
6643186
6643188
6643192
6643196
6643198
6643200
6643202
6643204
6643210
6643214
6643218
6643220
6643224
12666962
12830380
12830382
12830384
13276707
13877276
14279402
14279404
14279406
17226627
17226649
18307305
18307307
18307309
18307311
18307313
18307315
18307317
18307319
18307321
18307329
21311290
21311292
21669150
21669152
21669154
21669156
21669158
21669160
21669162
21669164
21669166
21669172
21669174
21669176
21669178
21669180
21669182
21669184
21669186
21669188
21669190
21669192
21669194
21669196
21669198
21669200
21669204
21669206
21669210
21669212
21669214
21669218
70888079
70888081
70888083
70888085
70888087
70888089
70888091
70888093
70888095
70888097
70888099
70888103
70888105
70888109
70888111
70888113
70888115
70888117
70888121
70888123
70888125
70888127
70888129
70888133
70888137
70888139
70888141
70888143
70888147
70888149
70888151
70888155
70888157
70888159
70888161
70888163
70888165
70888167
70888169
70888171
70888173
70888179
70888181
70888183
70888185
70888187
70888193
70888195
70888197
70888199
70888201
70888204
70888206
86439015
86439017
86439087
86439089
86439091
86439093
86439095
86439097
86439099
86439101
86439105
86439127
86439133
86439137
86439139
86439141
90994749
95007506
95007546
95007548
95007550
95007552
95007554
95007556
95007558
95007560
95007562
95007564
95007566
95007570
95007572
95007576
95007578
109240683
109240697
109240743
109240749
109240754
109240756
109240758
116795127
116795192
146336934
156632919
156632943
156632945
156632975
156633095
156633103
156633141
156633153
156633155
156633159

2078365
2654039
2654043
2865485
3023094
3023096
3023098
3023100
3023102
3023104
3023106
3023108
3023110
3023112
3023114
3023116
3023118
3023120
3023122
3023126
3023130
3023132
3091153
3091155
3091157
3091159
3091161
3091163
3091165
3091167
3091169
3091171
3091173
3091175
3091177
3091179
3091181
3091183
3091185
3091187
3091191
3091193
3091195
3091197
3091201
3091203
3091205
3091207
3091209
3091213
3093861
3093863
3093865
4103656
4103658
4103660
4103672
4324023
4324025
4324029
4324031
4324037
4324039
4324043
4324047
4324055
4324057
4324061
4324063
4324067
4324069
4324073
4324075
4324077
4324085
4324087
4324089
4324091
4324093
4324097
4324103
4324107
4324111
4324113
4324115
4324117
4324123
4324125
4324127
4324139
4324145
4324151
4324155
4324157
4324159
4324163
4324169
4324175
4324177
4324181
4324187
4324189
4324193
4324197
4324199
4324205
6643226
6643230
6643232
6643238
6643240
6643242
6643244
6643248
6643250
6643254
6643256
6643258
6643268
6643272
6643274
6643276
6643278
6643280
6643282
6643286
6643290
6643292
6643294
6643296
6643302
6643304
6643308
6643314
6643318
6643328
6643344
6643352
6643354
6643358
6643360
6643362
6643366
6643368
6643374
6643376
6643378
6643382
6643386
6643390
6643392
6643402
6643416
6643418
6643424
6643428
6643436
6643446
6643448
21669220
21669222
21669224
21669226
21669228
21669230
21669232
21669234
21669236
21669238
21669240
21669242
21669244
21669248
21669252
21669254
21669256
21669260
21669262
21669264
21669266
21669268
21669270
21669272
21669274
21669276
21669278
21669280
21669288
21998780
21998782
21998784
21998786
21998792
21998794
21998800
21998802
21998804
23194484
23194488
23194492
23194496
23343556
24474079
27369031
27369033
27369035
27369037
27369045
27369047
27369051
27369053
27369058
70888208
70888210
70888212
70888216
70888218
70888220
70888222
70888224
70888228
70888230
70888232
70888234
70888236
70888238
70888240
70888242
70888244
70888246
70888248
70888250
70888252
70888254
70888258
70888260
70888262
70888264
70888266
70888268
70888270
70888272
70888274
70888276
70888278
70888280
70888282
70888284
70888286
70888288
70888290
70888292
70888294
70888296
70888304
70888306
71482628
71482632
77378177
77378188
77378257
77378266
77378268
77378270
77378273
156633171
156633179
156633199
156633203
156633209
156633211
156633225
156633229
156633237
156633241
156633245
156633253
156633255
156633267
156633283
157093725
170684323
170684325
170684329
170684331
170684333
170684335
170684339
170684341
170684345
170684349
170684351
170684355
170684363
170684365
170684369
170684371
170684373
170684375
170684379
170684381
170684385
170684387
170684389
170684397
170684405
170684407
170684409
170684411
170684417
170684419
170684423
170684425
170684427
170684429
170684431
170684433
170684439

3093867
3093869
3093871
3093873
3093875
3093877
3093879
3093881
3093883
3093885
3093887
3093889
3093891
3093895
3093903
3142451
3142453
3142455
3142457
3142459
3142461
3142465
3142467
3142471
3142475
3142477
3142479
3142481
3142483
3142485
3142487
3142489
3142491
3142493
3142495
3142497
3142499
3142503
3142505
3142507
3142509
3142511
3142515
3142517
3142519
3142521
3142527
4324207
4324209
4324211
4324213
4324215
4324221
4324223
4324229
4324231
4324245
4324247
4324249
4324251
4324255
4324257
4324261
4324263
4324265
4324271
4324273
4324275
4324283
4324285
4468355
4468367
4468369
4468371
4565964
4565966
4565996
4566007
4566009
4566016
4566021
4566023
4566025
4566029
4566045
4566049
4566051
4566053
4566055
4566057
4566059
4566061
4566065
4566074
6643450
6643452
6643456
6643470
6643474
6643478
6643484
6643488
6643492
6643500
6643512
6643514
6643528
6643534
6643558
6643560
6643562
6643564
6643572
6643574
6643580
6643582
6643584
6643586
6643588
6643592
6643596
6643598
6643600
6643602
6643604
6643606
6643614
6643628
6643630
6649891
6649893
8920222
8920226
9864840
9968383
9968385
9968387
9968389
9968391
9968393
9968395
27369060
27369064
27369068
27369075
27369082
27369084
27369088
27818828
28394695
28394699
28394703
28394707
28394711
28394715
28848877
28848881
28848885
29342115
33304654
40647151
47271301
47271303
47271319
47271321
47271323
47271325
50199320
50199322
50199328
50199330
50199338
50199340
50871689
51103426
51103428
51103430
51103434
51103436
51103572
51103574
51103576
51103588
51103590
51103592
51103600
51103602
51103606
77378277
77378280
77378282
77378284
77378286
77378288
77378291
77378293
77378298
77378300
77378303
77378305
77378307
77378309
77378312
77378316
77378318
77378320
77378322
77378377
77378379
77378381
77378383
77378385
77378387
77378389
77378392
77378394
77378396
77378398
77378400
77378402
77379590
77379620
77379622
77379624
77379632
77379642
77379644
77379646
77379675
77379677
77379726
77379728
77379730
77379738
77379740
170684443
170684449
170684451
170684453
170684461
170684469
170684473
170684489
170684495
170684497
170684499
170684501
170684507
170684513
170684515
170684517
170684527
170684531
170684535
170684537
170684539
170684541
170684545
170684549
170684553
170684555
170684557
170684561
170684565
170684567
170684569
170684571
170684583
170684589
170684591
170684593
170684597
170684599
170684601
170684603
170684607
170684609
170684613
170684617
170684619

WHAT IS CLAIMED IS:

1. A library of synthetic polynucleotides, wherein said polynucleotides encode at least 10^6 unique antibody CDRH3 amino acid sequences comprising:

    (i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

    (ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

    (iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N2 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

    (iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

2. An antibody isolated from the polypeptide expression products of the library of claim 1.

3. The library of claim 1, wherein one or more CDRH3 amino acid sequences further comprise an N-terminal tail residue.

4. The library of claim 3, wherein the N-terminal tail residue is selected from the group consisting of G, D, and E.

5. The library of claim 1, wherein the N1 amino acid sequence is selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS,

KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof.

6. The library of claim 1, wherein the N2 amino acid sequence is selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof.

7. The library of claim 1, wherein the H3-JH amino acid sequence is selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV.

8. The library of claim 1, wherein the polynucleotides further comprise a 5' polynucleotide sequence and a 3' polynucleotide sequence that facilitate homologous recombination.

9. The library of claim 1, wherein the polynucleotides further encode one or more heavy chain chassis amino acid sequences that are N-terminal to the CDRH3 amino acid sequences, and the one or more heavy chain chassis amino acid sequences are selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 94 encoded by IGHV1-2, IGHV1-3, IGHV1-8, IGHV1-18, IGHV1-24, IGHV1-45, IGHV1-46, IGHV1-58, IGHV1-69, IGHV2-5, IGHV2-26, IGHV2-70, IGHV3-7, IGHV3-9, IGHV3-11, IGHV3-13, IGHV3-15, IGHV3-20, IGHV3-21, IGHV3-23, IGHV3-30, IGHV3-33, IGHV3-43, IGHV3-48, IGHV3-49, IGHV3-53, IGHV3-64, IGHV3-66, IGHV3-72, IGHV3-73, IGHV3-74, IGHV4-4, IGHV4-28, IGHV4-31, IGHV4-34, IGHV4-39, IGHV4-

59, IGHV4-61, IGHV4-B, IGHV5-51, IGHV6-1, and IGHV7-4-1, or a sequence of at least about 80% identity to any of them.

10. The library of claim 1, wherein the polynucleotides further encode one or more FRM4 amino acid sequences that are C-terminal to the CDRH3 amino acid sequences, wherein the one or more FRM4 amino acid sequences are selected from the group consisting of a FRM4 amino acid sequence encoded by IGHJ1, IGHJ2, IGHJ3, IGHJ4, IGHJ5, and IGHJ6, or a sequence of at least about 80% identity to any of them.

11. The library of claim 10, wherein the polynucleotides further encode one or more immunoglobulin heavy chain constant region amino acid sequences that are C-terminal to the FRM4 amino acid sequence.

12. The library of claim 11, wherein the CDRH3 amino acid sequences are expressed as part of full-length heavy chains.

13. The library of claim 12, wherein the full-length heavy chains are selected from the group consisting of an IgG1, IgG2, IgG3, and IgG4, or combinations thereof.

14. The library of claim 1, wherein the polynucleotides further encode an alternative scaffold.

15. The library of claim 1, wherein the CDRH3 amino acid sequences are from about 2 to about 30, from about 8 to about 19, or from about 10 to about 18 amino acid residues in length.

16. The library of claim 1, wherein the synthetic polynucleotides of the library encode from about 10^6 to about 10^{14}, from about 10^7 to about 10^{13}, from about 10^8 to about 10^{12}, from about 10^9 to about 10^{12}, or from about 10^{10} to about 10^{12} unique CDRH3 amino acid sequences.

17. A library of polypeptides encoded by the synthetic polynucleotide library of claim 1.

18. A library of vectors comprising the polynucleotide library of claim 1.

19. A population of cells comprising the vectors of claim 18.

20. The population of cells of claim 19, wherein the doubling time of the population of cells is from about 1 to about 3 hours, from about 3 to about 8 hours, from

about 8 to about 16 hours, from about 16 to about 20 hours, or from 20 to about 30 hours.

21. The population of cells of claim 19, wherein the cells are yeast cells.

22. The yeast cells of claim 21, wherein the yeast is *Saccharomyces cerevisiae*.

23. A library of synthetic polynucleotides, wherein said library has a theoretical total diversity of N unique CDRH3 sequences, wherein N is about 10^6 to about 10^{15}; and wherein the physical realization of the theoretical total CDRH3 diversity has a size of at least about 3N, thereby providing a probability of at least about 95% that any individual CDRH3 sequence contained within the theoretical total diversity of the library is present in the actual library.

24. A library of synthetic polynucleotides, wherein said polynucleotides encode at least about 10^6 unique antibody CDRH3 amino acid sequences comprising:

  (i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein:

    (a) the most N-terminal N1 amino acid, if present, is selected from a group consisting of R, G, P, L, S, A, V, K, I, Q, T and D;

    (b) the second most N-terminal N1 amino acid, if present, is selected from a group consisting of G, P, R, S, L, V, E, A, D, I, T and K; and

    (c) the third most N-terminal N1 amino acid, if present, is selected from the group consisting of G, R, P, S, L, A, V, T, E, D, K and F;

  (ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

  (iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein:

    (a) the most N-terminal N2 amino acid, if present, is selected from a group consisting of G, P, R, L, S, A, T, V, E, D, F and H;

(b) the second most N-terminal N2 amino acid, if present, is selected from a group consisting of G, P, R, S, T, L, A, V, E, Y, D and K; and

(c) the third most N-terminal N2 amino acid, if present, is selected from the group consisting of G, P, S, R, L, A, T, V, D, E, W and Q; and

(iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

25. An antibody isolated from the polypeptide expression products of the library of claim 24.

26. A library of synthetic polynucleotides, wherein said polynucleotides encode at least about 10^6 unique antibody CDRH3 amino acid sequences that are at least about 80% identical to an amino acid sequence represented by the following formula:

[X]-[N1]-[DH]-[N2]-[H3-JH], wherein:

(i) X is any amino acid residue or no amino acid residue;

(ii) N1 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof;

(iii) DH is an amino acid sequence selected from the group consisting of all possible reading frames that do not include a stop codon encoded by IGHD1-1, IGHD1-20, IGHD1-26, IGHD1-7, IGHD2-15, IGHD2-2, IGHD2-21, IGHD2-8, IGHD3-10, IGHD3-16, IGHD3-22, IGHD3-3, IGHD3-9, IGHD4-17, IGHD4-23, IGHD4-4, IGHD-4-11, IGHD5-12, IGHD5-24, IGHD5-5, IGHD-5-18, IGHD6-13, IGHD6-19, IGHD6-25, IGHD6-6, and IGHD7-27, and N- and C-terminal truncations thereof;

(iv) N2 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof; and

(v) H3-JH is an amino acid sequence selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV, or a sequence of at least 80% identity to any of them.

27. An antibody isolated from the polypeptide expression products of the library of claim 26.

28. A library of synthetic polynucleotides, wherein said library consists essentially of a plurality of polynucleotides encoding CDRH3 amino acid sequences that are at least about 80% identical to an amino acid sequence represented by the following formula:

[X]-[N1]-[DH]-[N2]-[H3-JH], wherein:

(i) X is any amino acid residue or no amino acid residue;

(ii) N1 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof;

(iii) DH is an amino acid sequence selected from the group consisting of all possible reading frames that do not include a stop codon encoded by IGHD1-1, IGHD1-20, IGHD1-26, IGHD1-7, IGHD2-15, IGHD2-2, IGHD2-21, IGHD2-8, IGHD3-10, IGHD3-16, IGHD3-22, IGHD3-3, IGHD3-9, IGHD4-17, IGHD4-23, IGHD4-4, IGHD-4-11, IGHD5-12, IGHD5-24, IGHD5-5, IGHD-5-18, IGHD6-13, IGHD6-19, IGHD6-25, IGHD6-6, and IGHD7-27, and N- and C-terminal truncations thereof;

(iv) N2 is an amino acid sequence selected from the group consisting of G, P, R, A, S, L, T, V, GG, GP, GR, GA, GS, GL, GT, GV, PG, RG, AG, SG, LG, TG, VG, PP, PR, PA, PS, PL, PT, PV, RP, AP, SP, LP, TP, VP, GGG, GPG, GRG, GAG, GSG, GLG, GTG, GVG, PGG, RGG, AGG, SGG, LGG, TGG, VGG, GGP, GGR, GGA, GGS, GGL, GGT, GGV, D, E, F, H, I, K, M, Q, W, Y, AR, AS, AT, AY, DL, DT, EA, EK, FH, FS, HL, HW, IS, KV, LD, LE, LR, LS, LT, NR, NT, QE, QL, QT, RA, RD, RE, RF, RH, RL, RR, RS, RV, SA, SD, SE, SF, SI, SK, SL, SQ, SR, SS, ST, SV, TA, TR, TS, TT, TW, VD, VS, WS, YS, AAE, AYH, DTL, EKR, ISR, NTP, PKS, PRP, PTA, PTQ, REL, RPL, SAA, SAL, SGL, SSE, TGL, WGT, and combinations thereof; and

(v) H3-JH is an amino acid sequence selected from the group consisting of AEYFQH, EYFQH, YFQH, FQH, QH, H, YWYFDL, WYFDL, YFDL, FDL, DL, L, AFDV, FDV, DV, V, YFDY, FDY, DY, Y, NWFDS, WFDS, FDS, DS, S, YYYYYGMDV, YYYYGMDV, YYYGMDV, YYGMDV, YGMDV, GMDV, and MDV, or a sequence of at least 80% identity to any of them.

29. An antibody isolated from the polypeptide expression products of the library of claim 28.

30. A library of synthetic polynucleotides, wherein said polynucleotides encode one or more antibody heavy chain amino acid sequences, and wherein the unique CDRH3 amino acid sequences of the heavy chain comprise:

(i) an N1 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

(ii) a human CDRH3 DH amino acid sequence, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(iii) an N2 amino acid sequence of 0 to about 3 amino acids, wherein each amino acid of the N2 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

(iv) a human CDRH3 H3-JH amino acid sequence, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them.

31. An antibody isolated from the polypeptide expression products of the library of claim 30.

32. A library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody VKCDR3 amino acid sequences comprising about 1 to about 10 of the amino acids found at Kabat positions 89, 90, 91, 92, 93, 94, 95, 95A, 96, and 97, in selected VKCDR3 amino acid sequences derived from a particular IGKV or IGKJ germline sequence.

33. The library of claim 32, wherein the synthetic polynucleotides encode one or more of the amino acid sequences listed in Table 33 or a sequence at least about 80% identical to any of them.

34. An antibody isolated from the polypeptide expression products of the library of claim 32.

35. A library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of unique antibody VKCDR3 amino acid sequences that are of at least about 80% identity to an amino acid sequence represented by the following formula:

[VK_Chassis]-[L3-VK]-[X]-[JK*], wherein:

(i) VK_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27, IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30, IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-21, and IGKV6D-41, or a sequence of at least about 80% identity to any of them;

(ii) L3-VK is the portion of the VKCDR3 encoded by the IGKV gene segment; and

(iii) X is any amino acid residue; and

(iv) JK* is an amino acid sequence selected from the group consisting of amino acid sequences encoded by IGJK1, IGJK2, IGJK3, IGJK4, and IGJK5, wherein the first amino acid residue of each amino acid sequence is not present.

36. An antibody isolated from the polypeptide expression products of the library of claim 35.

37. The library of claim 35, wherein X is selected from the group consisting of F, L, I, R, W, Y, and P.

38. A library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of VλCDR3 amino acid sequences that are of at least about 80% identity to an amino acid sequence represented by the following formula:

[Vλ_Chassis]-[L3-Vλ]-[Jλ], wherein:

(i) Vλ_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGλV1-36, IGλV1-40, IGλV1-44, IGλV1-47, IGλV1-51, IGλV10-54, IGλV2-11, IGλV2-14, IGλV2-18, IGλV2-23, IGλV2-8, IGλV3-1, IGλV3-10, IGλV3-12, IGλV3-16, IGλV3-19, IGλV3-21, IGλV3-25, IGλV3-27, IGλV3-9, IGλV4-3, IGλV4-60, IGλV4-69, IGλV5-39, IGλV5-45, IGλV6-57, IGλV7-43, IGλV7-46, IGλV8-61, IGλV9-49, and IGλV10-54, or a sequence of at least about 80% identity to any of them;

(ii) L3-Vλ is the portion of the VλCDR3 encoded by the IGλV segment; and

(iii) Jλ is an amino acid sequence selected from the group consisting of amino acid sequences encoded by IGλJ1-01, IGλJ2-01, IGλJ3-01, IGλJ3-02, IGλJ6-01, IGλJ7-01, and IGλJ7-02, and wherein the first amino acid residue of each sequence may or may not be deleted.

39. An antibody isolated from the polypeptide expression products of the library of claim 38.

40. A library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody proteins comprising:

(i) a CDRH3 amino acid sequence of claim 1; and

(ii) a VKCDR3 amino acid sequence comprising about 1 to about 10 of the amino acids found at Kabat positions 89, 90, 91, 92, 93, 94, 95, 95A, 96, and 97, in selected VKCDR3 amino acid sequences derived from a particular IGKV or IGKJ germline sequence.

41. The library of claim 40, wherein the VKCDR3 amino acid sequence comprises one or more of the amino acid sequences listed in Table 33 or a sequence at least about 80% identical to any of them.

42. An antibody isolated from the library of claim 40.

43. The library of claim 40, wherein the antibody proteins are expressed in a heterodimeric form.

44. The library of claim 40, wherein the antibody proteins are expressed as antibody fragments.

45. The library of claim 44, wherein the antibody fragments are selected from the group consisting of Fab, Fab', F(ab')$_2$, Fv fragments, diabodies, linear antibodies, and single-chain antibodies.

46. A library of synthetic polynucleotides, wherein said polynucleotides encode a plurality of antibody proteins comprising:

(i) a CDRH3 amino acid sequence of claim 1; and

(ii) a VKCDR3 amino acid sequence of at least about 80% identity to a amino acid sequence represented by the following formula:

[VK_Chassis]-[L3-VK]-[X]-[JK*], wherein:

(a) VK_Chassis is an amino acid sequence selected from the group consisting of about Kabat amino acid 1 to about Kabat amino acid 88 encoded by IGKV1-05, IGKV1-06, IGKV1-08, IGKV1-09, IGKV1-12, IGKV1-13, IGKV1-16, IGKV1-17, IGKV1-27, IGKV1-33, IGKV1-37, IGKV1-39, IGKV1D-16, IGKV1D-17, IGKV1D-43, IGKV1D-8, IGKV2-24, IGKV2-28, IGKV2-29, IGKV2-30, IGKV2-40, IGKV2D-26, IGKV2D-29, IGKV2D-30,

IGKV3-11, IGKV3-15, IGKV3-20, IGKV3D-07, IGKV3D-11, IGKV3D-20, IGKV4-1, IGKV5-2, IGKV6-21, and IGKV6D-41, or a sequence of at least about 80% identity to any of them;

(b) L3-VK is the portion of the VKCDR3 encoded by the IGKV gene segment; and

(c) X is any amino acid residue; and

(d) JK* is an amino acid sequence selected from the group consisting of amino acid sequences encoded by IGJK1, IGJK2, IGJK3, IGJK4, and IGJK5, wherein the first residue of each IGJK amino acid sequence is not present.

47. An antibody isolated from the library of claim 46.

48. A library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+1 pairs in the library is within the ranges specified below:

Tyr-Tyr in an amount from about 2.5% to about 6.5%;

Ser-Gly in an amount from about 2.5% to about 4.5%;

Ser-Ser in an amount from about 2% to about 4%;

Gly-Ser in an amount from about 1.5% to about 4%;

Tyr-Ser in an amount from about 0.75% to about 2%;

Tyr-Gly in an amount from about 0.75% to about 2%; and

Ser-Tyr in an amount from about 0.75% to about 2%.

49. The library of claim 48, wherein at least 2, 3, 4, 5, 6, or 7 of the specified i – i+1 pairs in the library are within the specified ranges.

50. An antibody isolated from the polypeptide expression products of the library of claim 48.

51. The library of claim 48, wherein the polynucleotides encode at least about 10^6 unique CDRH3 amino acid sequences.

52. A library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+2 pairs in the library is within the ranges specified below:

Tyr-Tyr in an amount from about 2.5% to about 4.5%;

Gly-Tyr in an amount from about 2.5% to about 5.5%;

Ser-Tyr in an amount from about 2% to about 4%;

Tyr-Ser in an amount from about 1.75% to about 3.75%;

Ser-Gly in an amount from about 2% to about 3.5%;

Ser-Ser in an amount from about 1.5% to about 3%;

Gly-Ser in an amount from about 1.5% to about 3%; and

Tyr-Gly in an amount from about 1% to about 2%.

53. The library of claim 52, wherein at least 2, 3, 4, 5, 6, 7, or 8 of the specified i – i+2 pairs in the library are within the specified ranges.

54. A antibody isolated from the polypeptide expression products of the library of claim 52.

55. The library of claim 52, wherein the polynucleotides encode at least about 10^6 unique CDRH3 amino acid sequences.

56. A library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, wherein the percent occurrence within the central loop of the CDRH3 amino acid sequences of at least one of the following i – i+3 pairs in the library is within the ranges specified below:

Gly-Tyr in an amount from about 2.5% to about 6.5%;

Ser-Tyr in an amount from about 1% to about 5%;

Tyr-Ser in an amount from about 2% to about 4%;

Ser-Ser in an amount from about 1% to about 3%;

Gly-Ser in an amount from about 2% to about 5%; and

Tyr-Tyr in an amount from about 0.75% to about 2%.

57. The library of claim 56, wherein at least 2, 3, 4, 5, or 6 of the specified i – i+3 pairs in the library are within the specified ranges.

58. An antibody isolated from the polypeptide expression products of the library of claim 56.

59. The library of claim 56, wherein the polynucleotides encode at least about 10^6 unique CDRH3 amino acid sequences.

60. A method of preparing a synthetic polynucleotide library comprising providing and assembling the polynucleotide sequences of any one of claims 1, 24, 26, 28, 30, 32, 35, 38, 40, 46, 48, 52, and 56.

61. A method of preparing a library of synthetic polynucleotides encoding a plurality of antibody CDRH3 amino acid sequences, the method comprising:

(i) providing polynucleotide sequences encoding:

(a) one or more N1 amino acid sequences of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N1 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells;

(b) one or more human CDRH3 DH amino acid sequences, N- and C-terminal truncations thereof, or a sequence of at least about 80% identity to any of them;

(c) one or more N2 amino acid sequences of 0 to about 3 amino acids, wherein each amino acid of the N1 amino acid sequence is among the 12 most frequently occurring amino acids at the corresponding position in N2 amino acid sequences of CDRH3 amino acid sequences that are functionally expressed by human B cells; and

(d) one or more human CDRH3 H3-JH amino acid sequences, N-terminal truncations thereof, or a sequence of at least about 80% identity to any of them; and

(ii) assembling the polynucleotide sequences to produce a library of synthetic polynucleotides encoding a plurality of human antibody CDRH3 amino acid sequences represented by the following formula:

[N1]-[DH]-[N2]-[H3-JH].

62. The method of claim 61, wherein one or more of the polynucleotide sequences are synthesized via split-pool synthesis.

63. The method of claim 61, further comprising the step of providing a 5' polynucleotide sequence and a 3' polynucleotide sequence that facilitate homologous recombination.

64. The method of claim 61, further comprising the step of recombining the assembled synthetic polynucleotides with a vector comprising a heavy chain chassis and a heavy chain constant region, to form a full-length heavy chain.

65. The method of claim 64, wherein the step of recombining is performed in yeast.

66. The method of claim 65, wherein the yeast is *S. cerevisiae*.

67. A method of isolating one or more host cells expressing one or more antibodies, the method comprising:

(i) expressing the antibodies of any one of claims 40 and 46 in one or more host cells,

(ii) contacting the host cells with one or more antigens; and

(iii) isolating one or more host cells having antibodies that bind to the one or more antigens.

68. The method of claim 67, further comprising isolating one or more antibodies from the one or more host cells.

69. The method of claim 67, further comprising the step of isolating one or more polynucleotide sequences encoding one or more antibodies from the one or more host cells.

70. A kit comprising the library of synthetic polynucleotides of claim 1.

71. The CDRH3 amino acid sequences encoded by the library of synthetic polynucleotides of any one of claims 1, 24, 26, 28, 30, 32, 35, 38, 40, 46, 48, 52, and 56 in computer readable form.

Figure 1

Construction by yeast homologous recombination

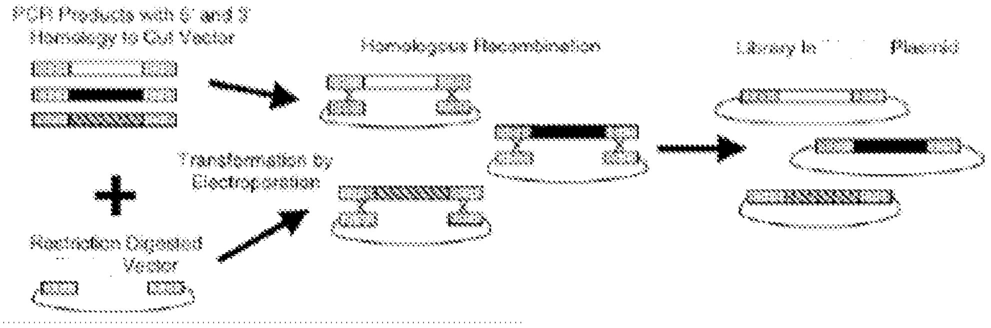


Figure 2

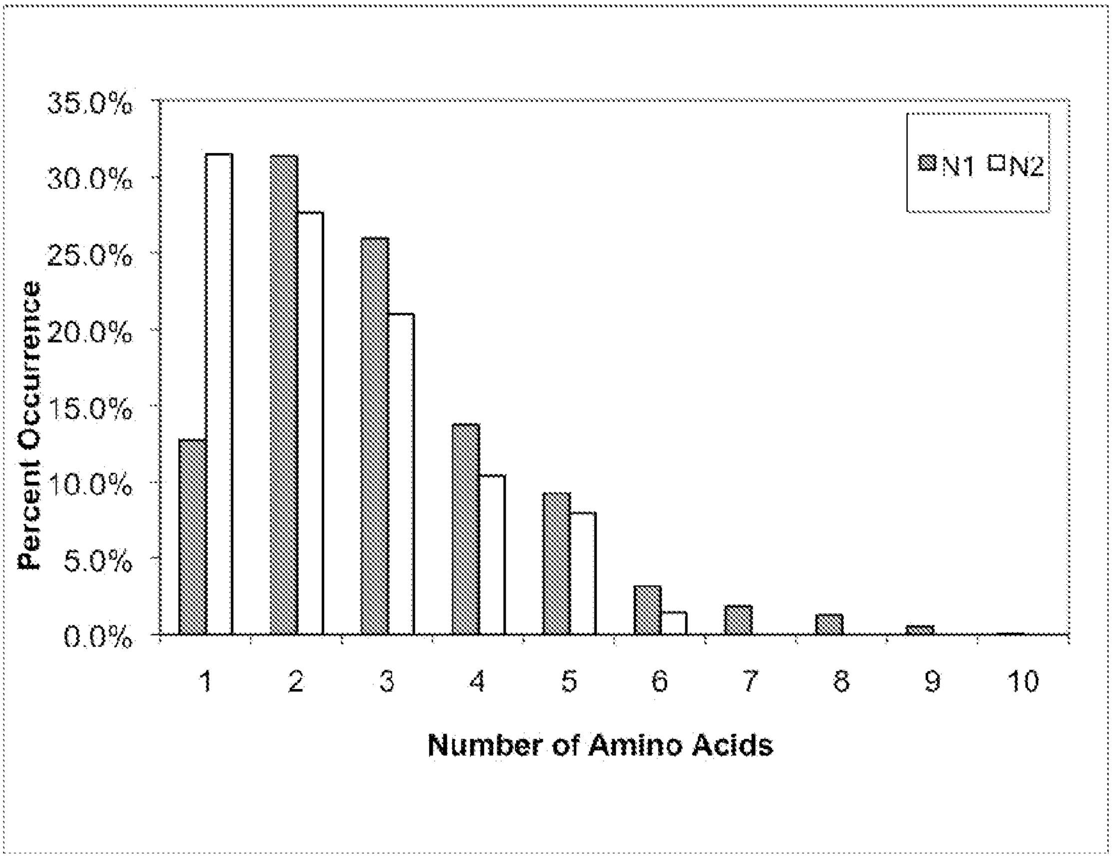

35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Percent Occurrence
▨N1 □N2
1
2
3
4
5
6
7
8
9
10
Number of Amino Acids

Figure 3

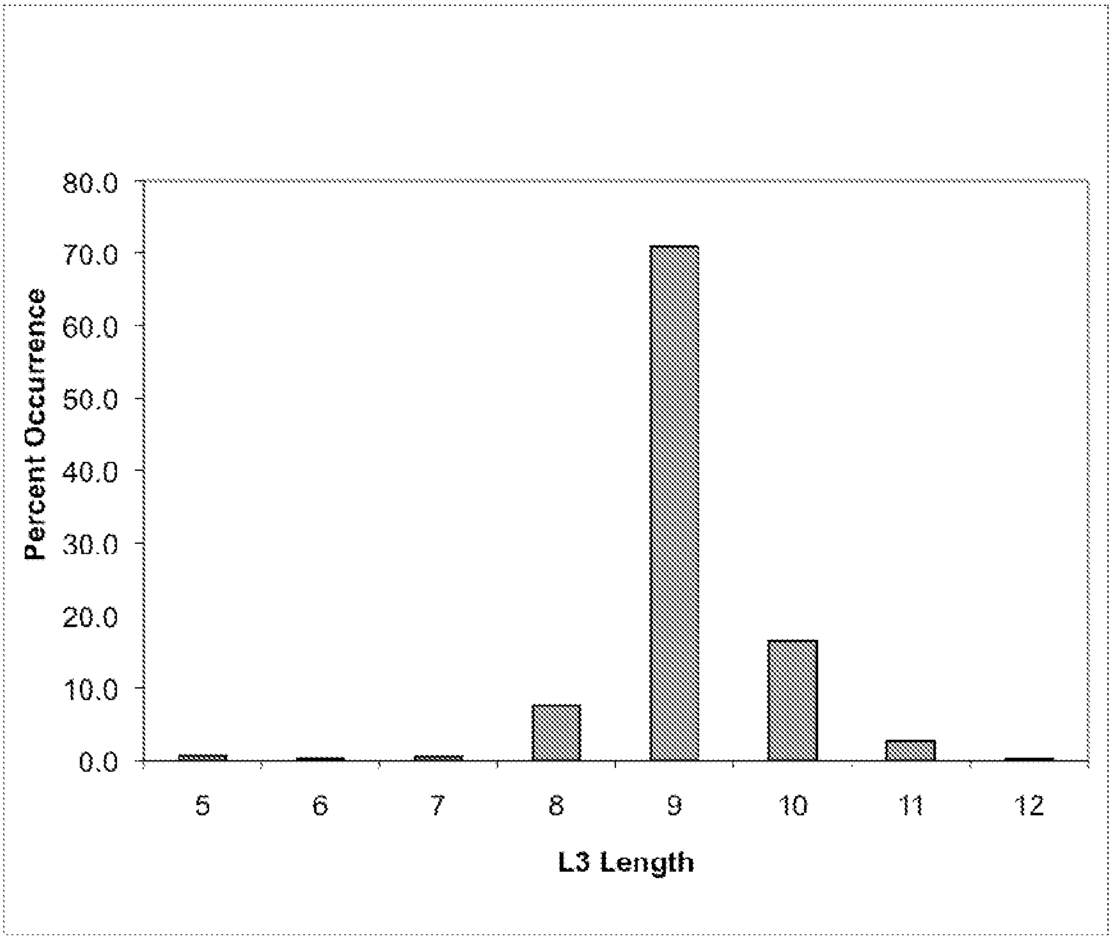

80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Percent Occurrence
5
6
7
8
9
10
11
12
L3 Length

Figure 4

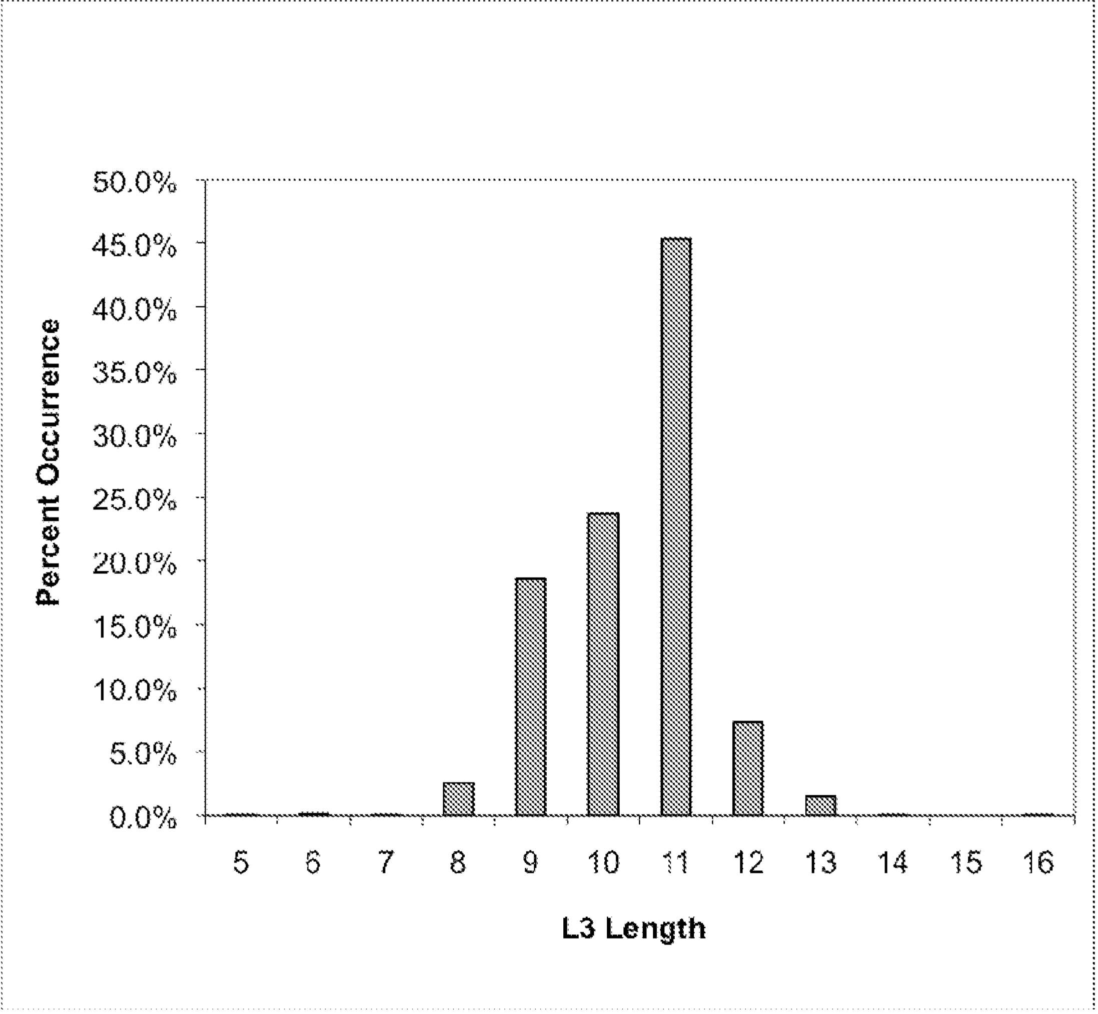

50.0%
45.0%
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Percent Occurrence
5
6
7
8
9
10
11
12
13
14
15
16
L3 Length

Figure 5

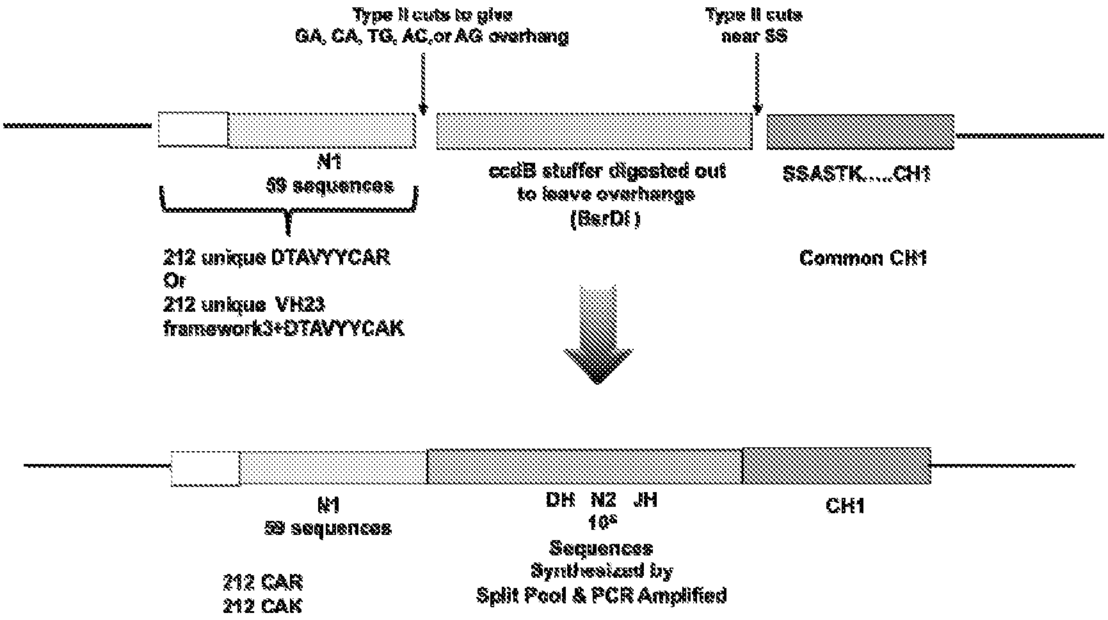

Type II cuts to give
GA, CA, TG, AC, or AG overhang
Type II cuts
near SS
N1
59 sequences
ccdB stuffer digested out
to leave overhang
(BsrDI)
SSASTK.....CH1
212 unique DTAVYYCAR
Or
212 unique VH23
framework3+DTAVYYCAK
Common CH1
N1
59 sequences
DH N2 JH
10^6
Sequences
Synthesized by
Split Pool & PCR Amplified
CH1
212 CAR
212 CAK

Figure 6

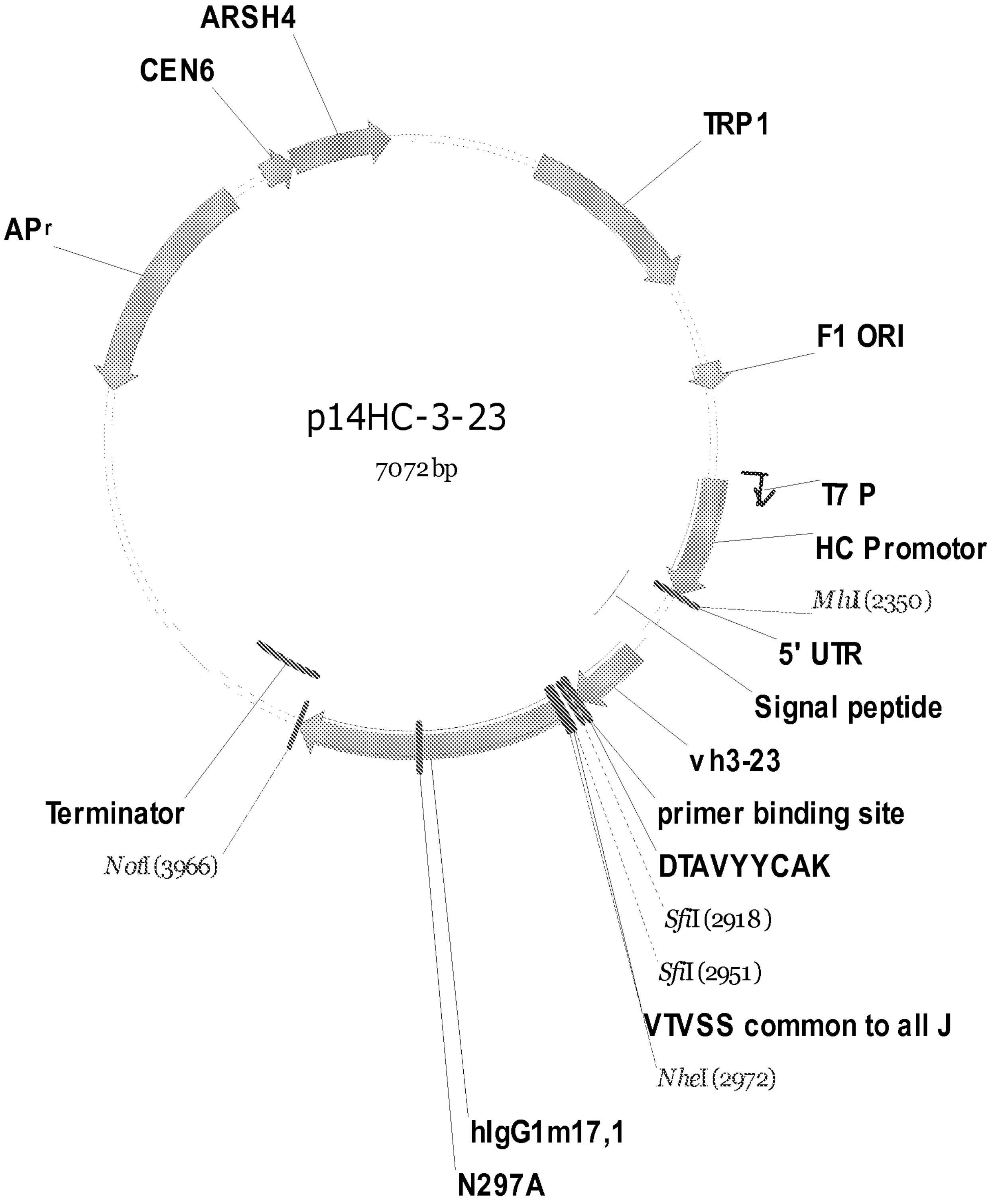
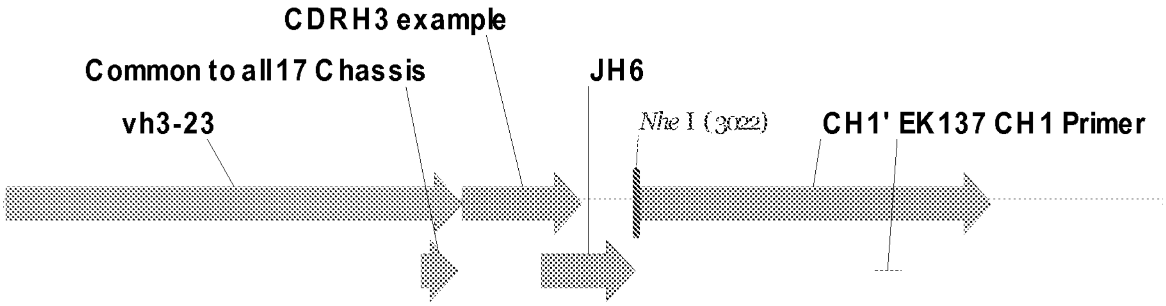

ARSH4
CEN6
TRP1
AP^r
F1 ORI
p14HC-3-23
7072 bp
T7 P
HC Promotor
MluI (2350)
5' UTR
Signal peptide
vh3-23
primer binding site
DTAVYYCAK
SfiI (2918)
SfiI (2951)
VTVSS common to all J
NheI (2972)
Terminator
NotI (3966)
hIgG1m17,1
N297A

Figure 7



Fragment of p14HC-3-23

745 bp (molecule 7122 bp)

```
                                         vh3-23
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         E  V  Q  L   L  E  S   G  G  G   L  V  Q  P   G  G  S   L  R  L   S  C  A  A   S  G  F ·
2616    GAGGTGCAGC TGTTGGAGTC TGGGGGAGGC TTGGTACAGC CTGGGGGGTC CCTGAGACTC TCCTGTGCAG CCTCTGGATT
        CTCCACGTCG ACAACCTCAG ACCCCCTCCG AACCATGTCG GACCCCCCAG GGACTCTGAG AGGACACGTC GGAGACCTAA
                                         vh3-23
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        · T  F  S   S  Y  A  M   S  W  V   R  Q  A   P  G  K  G   L  E  W   V  S  A   I  S  G  S ·
2696    CACCTTTAGC AGCTATGCCA TGAGCTGGGT CCGCCAGGCT CCAGGGAAGG GGCTGGAGTG GGTCTCAGCT ATTAGTGGTA
        GTGGAAATCG TCGATACGGT ACTCGACCCA GGCGGTCCGA GGTCCCTTCC CCGACCTCAC CCAGAGTCGA TAATCACCAT
                                         vh3-23
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ·  G  G  S   T  Y  Y   A  D  S  V   K  G  R   F  T  I   S  R  D  N   S  K  N   T  L  Y
2776    GTGGTGGTAG CACATACTAC GCAGACTCCG TGAAGGGCCG GTTCACCATC TCCAGAGACA ATTCCAAGAA CACGCTGTAT
        CACCACCATC GTGTATGATG CGTCTGAGGC ACTTCCCGGC CAAGTGGTAG AGGTCTCTGT TAAGGTTCTT GTGCGACATA
                                  Common to all 17 Chassis         CDRH3 example
                                  ~~~~~~~~~~~~~~~~~~~~~~~~   ~~~~~~~~~~~~~~~~~~~~~~~~~~
                              vh3-23
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         L  Q  M  N   S  L  R   A  E  D   T  A  V  Y   Y  C  A   K  E  G   G  P  G  Y   C  S  S ·
2856    CTGCAAATGA ACAGCCTGAG AGCCGAGGAC ACGGCGGTGT ACTACTGCGC CAAGGAGGGC GGACCTGGGT ATTGCAGTTC
        GACGTTTACT TGTCGGACTC TCGGCTCCTG TGCCGCCACA TGATGACGCG GTTCCTCCCG CCTGGACCCA TAACGTCAAG
                                                        JH6
                                                      ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                         CDRH3 example
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        · T  S  C   Y  T  P  G   G  Y  Y   Y  Y  Y   G  M  D  V   W  G  Q   G  T  T   V  T  V  S ·
2936    AACTTCTTGT TATACACCAG GAGGCTACTA CTATTACTAC GGCATGGACG TGTGGGGACA AGGTACAACA GTCACCGTCT
        TTGAAGAACA ATATGTGGTC CTCCGATGAT GATAATGATG CCGTACCTGC ACACCCCTGT TCCATGTTGT CAGTGGCAGA
                                            CH1'
             ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         JH6
        ~~~~~
             NheI
             ~~~~~~~
        ·  S  A  S   T  K  G   P  S  V  F   P  L  A   P  S  S   K  S  T  S   G  G  T   A  A  L
3016    CCTCAGCTAG CACCAAGGGC CCATCGGTCT TCCCCCTGGC ACCCTCCTCC AAGAGCACCT CTGGGGGCAC AGCGGCCCTG
        GGAGTCGATC GTGGTTCCCG GGTAGCCAGA AGGGGGACCG TGGGAGGAGG TTCTCGTGGA GACCCCCGTG TCGCCGGGAC
                                          CH1'
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                                                                                      EK137 CH1
        Primer
                                                                                      ~
         G  C  L  V   K  D  Y   F  P  E   P  V  T  V   S  W  N   S  G  A   L  T  S  G   V  H  T ·
3096    GGCTGCCTGG TCAAGGACTA CTTCCCCGAA CCGGTGACGG TGTCGTGGAA CTCAGGCGCC CTGACCAGCG GCGTGCACAC
        CCGACGGACC AGTTCCTGAT GAAGGGGCTT GGCCACTGCC ACAGCACCTT GAGTCCGCGG GACTGGTCGC CGCACGTGTG
                                        CH1'
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        EK137 CH1 Primer
        ~~~~~~~~~~~~~~~~~
        · F  P  A   V  L  Q  S   S  G  L   Y  S  L   S  S  V  V   T  V  P   S  S  S   L
3176    CTTCCCGGCT GTCCTACAGT CCTCAGGACT CTACTCCCTC AGCAGCGTGG TGACCGTGCC CTCCAGCAGC TTG
        GAAGGGCCGA CAGGATGTCA GGAGTCCTGA GATGAGGGAG TCGTCGCACC ACTGGCACGG GAGGTCGTCG AAC
```

Figure 8

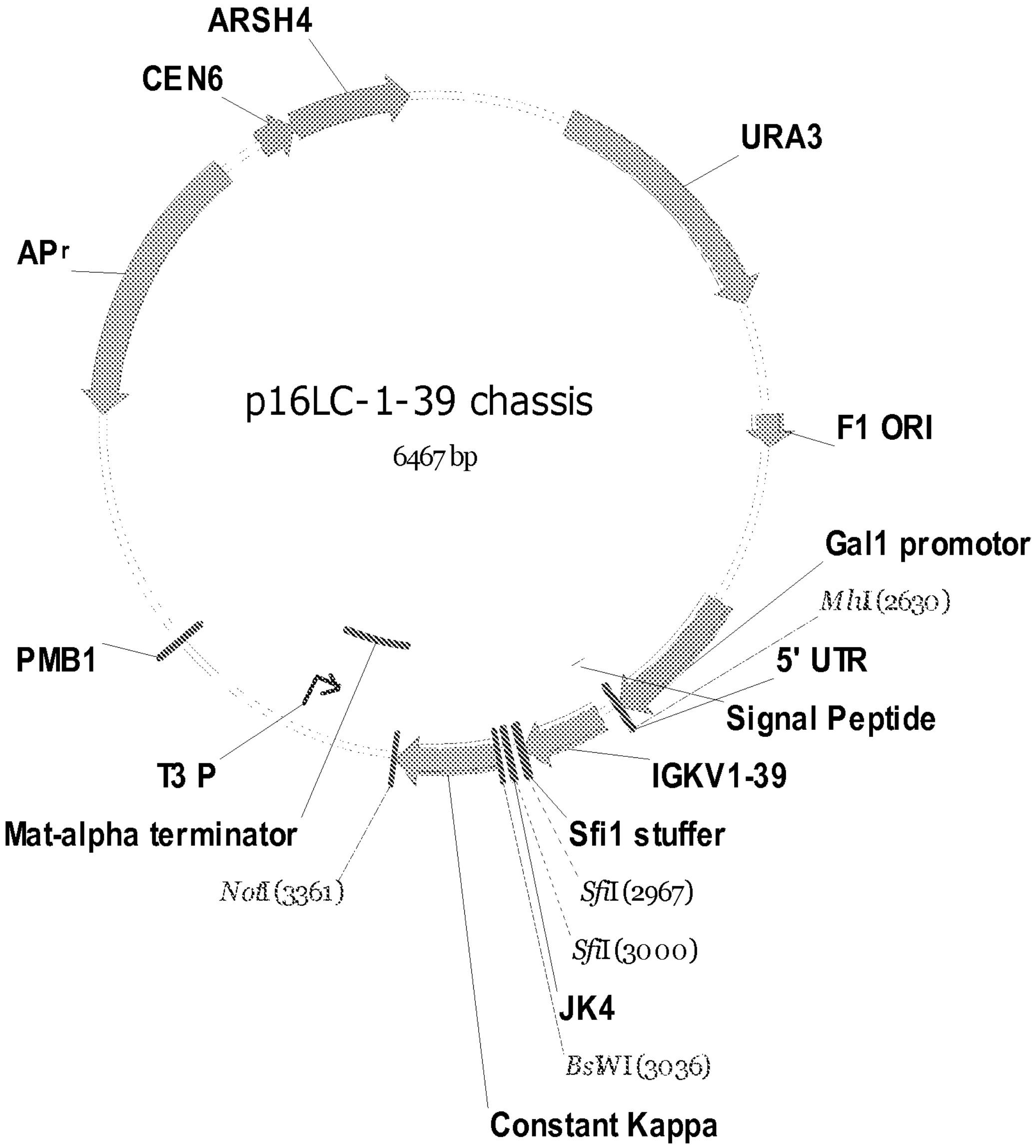
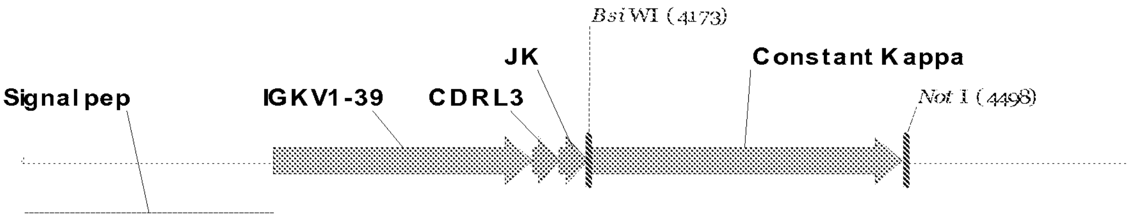

ARSH4
CEN6
URA3
APr
p16LC-1-39 chassis
6467 bp
F1 ORI
Gal1 promotor
MluI (2630)
5' UTR
Signal Peptide
PMB1
IGKV1-39
T3 P
Mat-alpha terminator
Sfi1 stuffer
NotI (3361)
SfiI (2967)
SfiI (3000)
JK4
BsiWI (3036)
Constant Kappa

Figure 9



Fragment of p15LC-1-39-LC3-9aa

1128 bp (molecule 7604 bp)

```
                                            Signal pep
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
           M  R  F   P  S  I   F  T  A  V   L  F  A   A  S  S   A  L  A  A   P  A  N   T  T  T  E ·
3595     AATGCGCTTT CCGAGCATTT TTACCGCAGT TCTGTTTGCG GCGAGCAGCG CGCTGGCGGC GCCGGCGAAC ACCACCACCG
         TTACGCGAAA GGCTCGTAAA AATGGCGTCA AGACAAACGC CGCTCGTCGC GCGACCGCCG CGGCCGCTTG TGGTGGTGGC
                                            Signal pep
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         ·  D  E  T   A  Q  I   P  A  E  A   V  I  D   Y  S  D   L  E  G  D   F  D  A   A  A  L
3675     AAGATGAAAC CGCGCAGATT CCGGCGGAAG CGGTGATTGA TTATAGCGAT CTGGAAGGCG ATTTTGATGC GGCGGCGCTG
         TTCTACTTTG GCGCGTCTAA GGCCGCCTTC GCCACTAACT AATATCGCTA GACCTTCCGC TAAAACTACG CCGCCGCGAC
                                            Signal pep
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          P  L  S  N   S  T  N   N  G  L   S  S  T  N   T  T  I   A  S  I   A  A  K  E   E  G  V ·
3755     CCGCTGAGCA ACAGCACCAA CAACGGCCTG AGCAGCACCA ACACCACCAT TGCGAGCATT GCGGCGAAAG AAGAAGGCGT
         GGCGACTCGT TGTCGTGGTT GTTGCCGGAC TCGTCGTGGT TGTGGTGGTA ACGCTCGTAA CGCCGCTTTC TTCTTCCGCA
            Signal pep
         ~~~~~~~~~~~~~~~~~
                                                      IGKV1-39
                          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         · Q  L  D   K  R  D  I   Q  M  T   Q  S  P   S  S  L  S   A  S  V   G  D  R   V  T  I  T ·
3835     GCAGCTGGAT AAACGCGACA TCCAGATGAC CCAGTCTCCA TCCTCCCTGT CTGCATCTGT AGGAGACAGA GTCACCATCA
         CGTCGACCTA TTTGCGCTGT AGGTCTACTG GGTCAGAGGT AGGAGGGACA GACGTAGACA TCCTCTGTCT CAGTGGTAGT
                                             IGKV1-39
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         ·  C  R  A   S  Q  S   I  S  S  Y   L  N  W   Y  Q  Q   K  P  G  K   A  P  K   L  L  I
3915     CTTGCCGGGC AAGTCAGAGC ATTAGCAGCT ATTTAAATTG GTATCAGCAG AAACCAGGGA AAGCCCCTAA GCTCCTGATC
         GAACGGCCCG TTCAGTCTCG TAATCGTCGA TAAATTTAAC CATAGTCGTC TTTGGTCCCT TTCGGGGATT CGAGGACTAG
                                             IGKV1-39
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          Y  A  A  S   S  L  Q   S  G  V   P  S  R  F   S  G  S   G  S  G   T  D  F  T   L  T  I ·
3995     TATGCTGCAT CCAGTTTGCA AAGTGGGGTC CCATCAAGGT TCAGTGGCAG TGGATCTGGG ACAGATTTCA CTCTCACCAT
         ATACGACGTA GGTCAAACGT TTCACCCCAG GGTAGTTCCA AGTCACCGTC ACCTAGACCC TGTCTAAAGT GAGAGTGGTA
                                                                CDRL3
                                                     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                         IGKV1-39                                                           JK
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~                             ~~~~~~~~~~~~~~
         · S  S  L   Q  P  E  D   F  A  T   Y  Y  C   Q  Q  S  Y   S  T  P   L  T  F   G  G  G  T ·
4075     CAGCAGTCTG CAACCTGAAG ATTTTGCAAC TTACTACTGT CAACAGAGTT ACAGTACCCC TCTCACTTTT GGCGGAGGGA
         GTCGTCAGAC GTTGGACTTC TAAAACGTTG AATGATGACA GTTGTCTCAA TGTCATGGGG AGAGTGAAAA CCGCCTCCCT
                JK
         ~~~~~~~~~~~~~~~~~~
                                                   Constant Kappa
                           ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
                           BsiWI
                           ~~~~~~~
         ·  K  V  E   I  K  R   T  V  A  A   P  S  V   F  I  F   P  P  S  D   E  Q  L   K  S  G
4155     CCAAGGTTGA GATCAAACGT ACGGTGGCCG CTCCTTCCGT GTTCATCTTC CCTCCCTCCG ACGAGCAGCT GAAGTCCGGC
         GGTTCCAACT CTAGTTTGCA TGCCACCGGC GAGGAAGGCA CAAGTAGAAG GGAGGGAGGC TGCTCGTCGA CTTCAGGCCG
                                          Constant Kappa
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
          T  A  S  V   V  C  L   L  N  N   F  Y  P  R   E  A  K   V  Q  W   K  V  D  N   A  L  Q ·
4235     ACCGCCAGCG TGGTGTGCCT GCTGAACAAC TTCTACCCTC GGGAGGCCAA GGTGCAGTGG AAGGTGGACA ACGCCCTGCA
         TGGCGGTCGC ACCACACGGA CGACTTGTTG AAGATGGGAG CCCTCCGGTT CCACGTCACC TTCCACCTGT TGCGGGACGT
                                          Constant Kappa
         ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
         · S  G  N   S  Q  E  S   V  T  E   Q  D  S   K  D  S  T   Y  S  L   S  S  T   L  T  L  S ·
4315     GAGCGGCAAC TCCCAGGAGT CCGTCACCGA GCAGGACTCC AAGGACAGCA CCTACTCCCT GTCCTCCACC CTGACCCTGT
         CTCGCCGTTG AGGGTCCTCA GGCAGTGGCT CGTCCTGAGG TTCCTGTCGT GGATGAGGGA CAGGAGGTGG GACTGGGACA
```

```
                                        Constant Kappa
        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        · K  A  D   Y  E  K   H  K  V  Y   A  C  E   V  T  H   Q  G  L  S   S  P  V   T  K  S
4395    CCAAGGCCGA CTACGAGAAG CATAAGGTGT ACGCCTGCGA GGTGACCCAC CAGGGCCTGT CCAGCCCTGT GACCAAGTCC
        GGTTCCGGCT GATGCTCTTC GTATTCCACA TGCGGACGCT CCACTGGGTG GTCCCGGACA GGTCGGGACA CTGGTTCAGG
          Constant Kappa
        ~~~~~~~~~~~~~~~~~~~
                                 NotI
                               ~~~~~~~~
         F  N  R  G   E  C  *
4475    TTCAACCGGG GCGAGTGCTA GGCGGCCGCA
        AAGTTGGCCC CGCTCACGAT CCGCCGGCGT
```

Figure 10

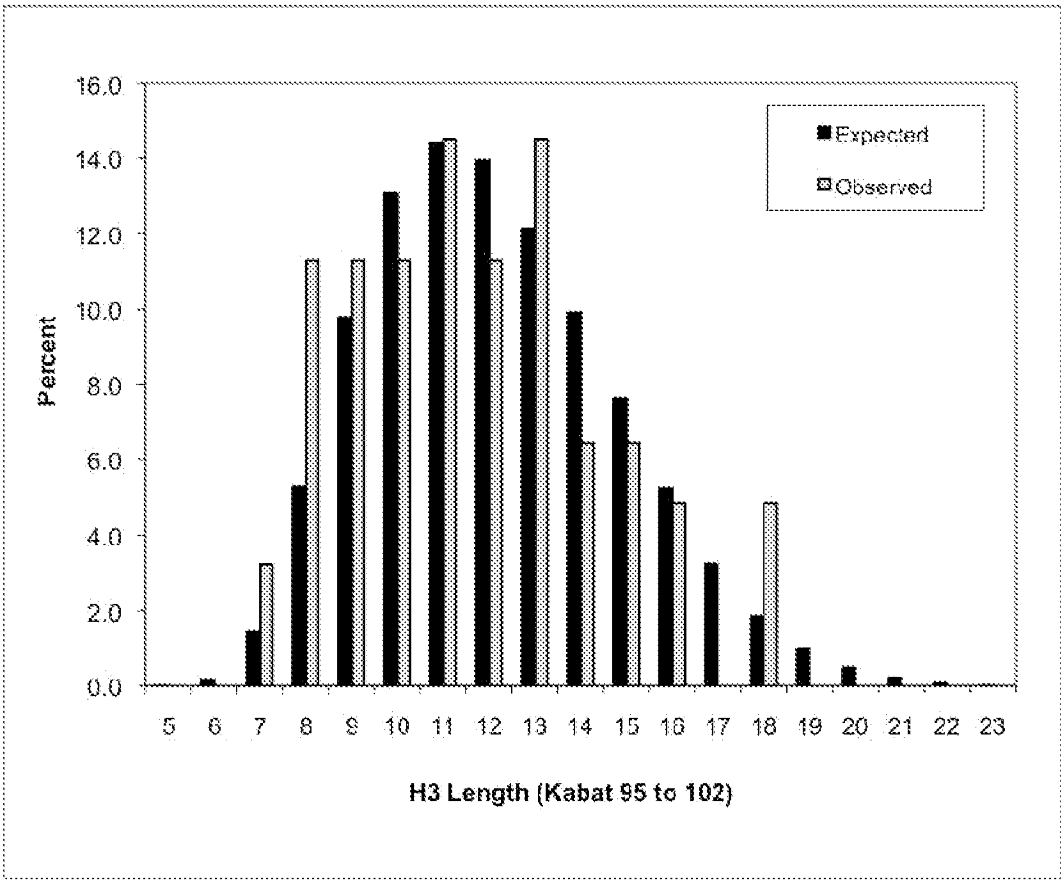

16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Percent
■ Expected
▨ Observed
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
H3 Length (Kabat 95 to 102)

Figure 11

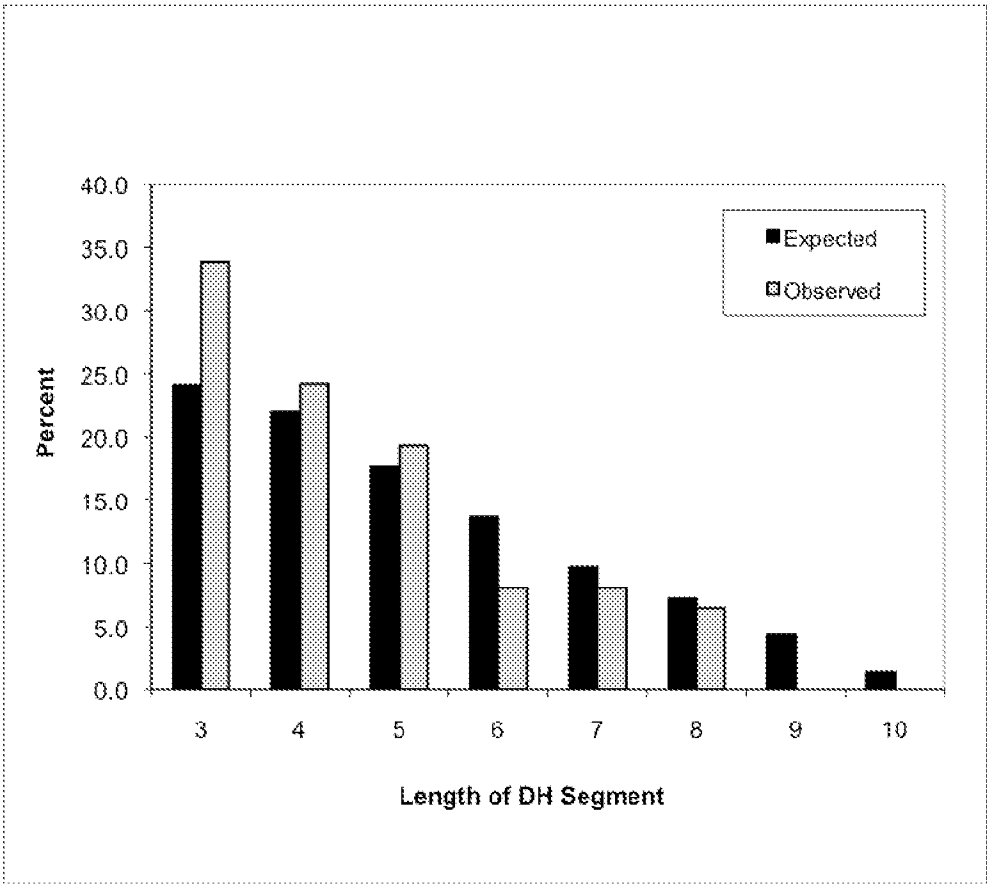

40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percent
■ Expected
▨ Observed
3
4
5
6
7
8
9
10
Length of DH Segment

Figure 12

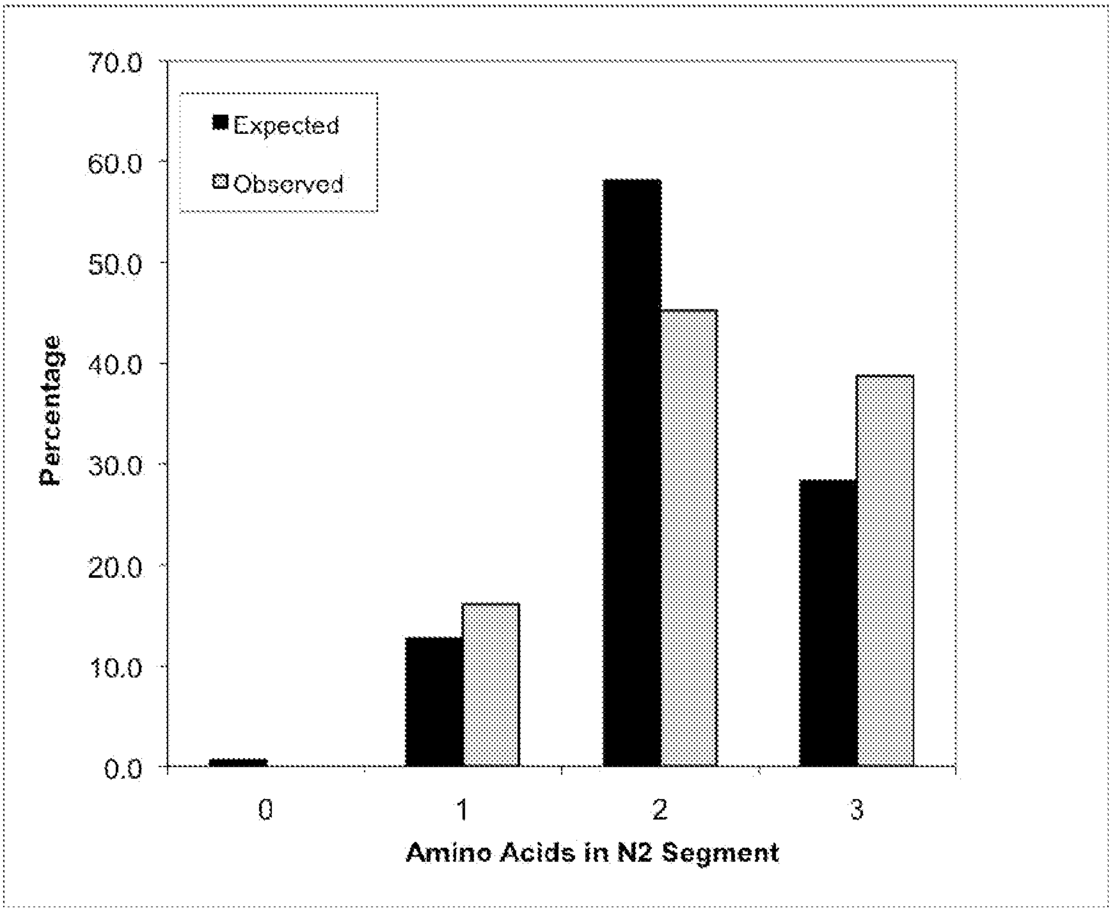

70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Percentage
Expected
Observed
0
1
2
3
Amino Acids in N2 Segment

Figure 13

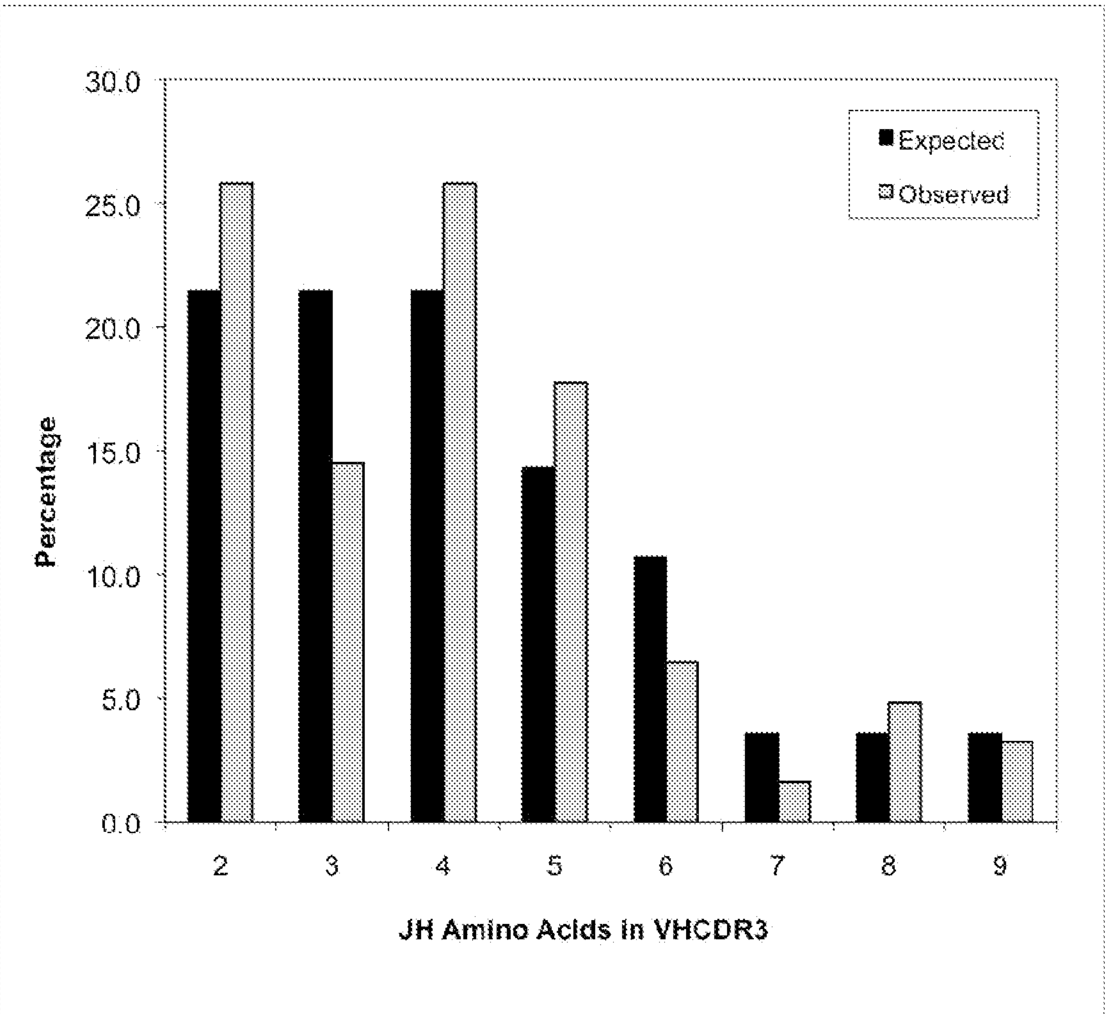

30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
■ Expected
▨ Observed
2
3
4
5
6
7
8
9
JH Amino Acids in VHCDR3

Figure 14

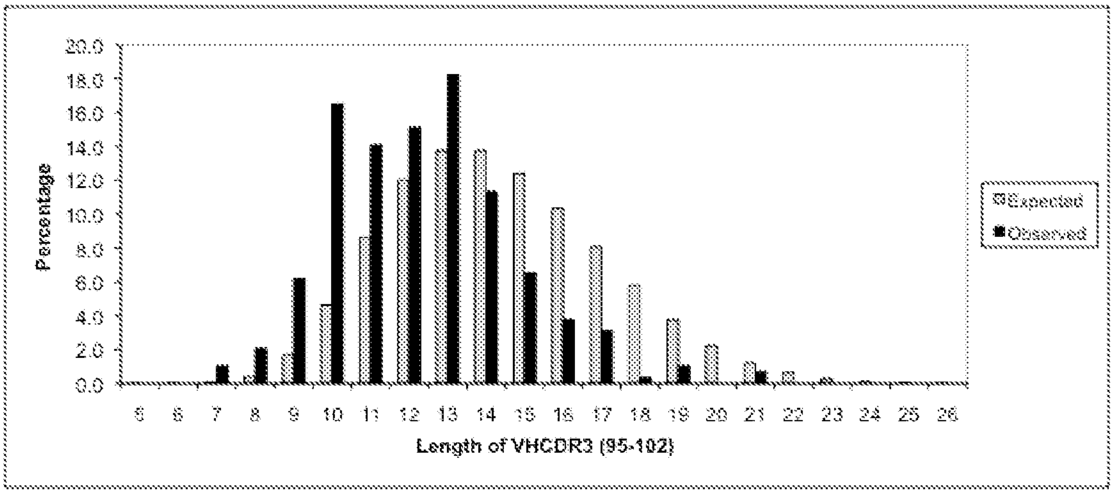

20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Percentage
Expected
Observed
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
Length of VHCDR3 (95-102)

Figure 15

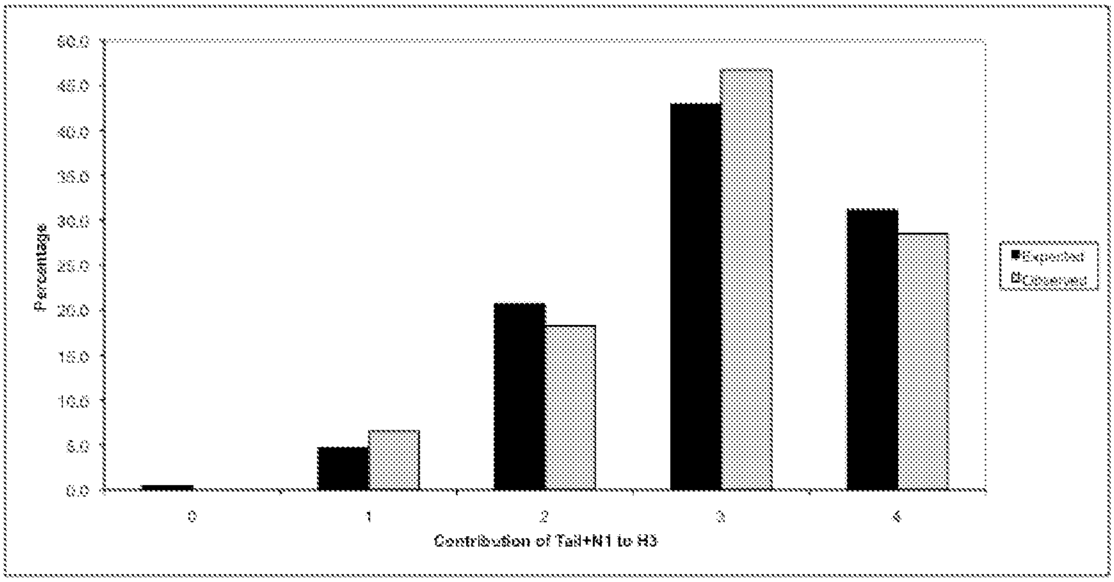

50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
0
1
2
3
4
Contribution of Tail+N1 to H3
Expected
Observed

Figure 16

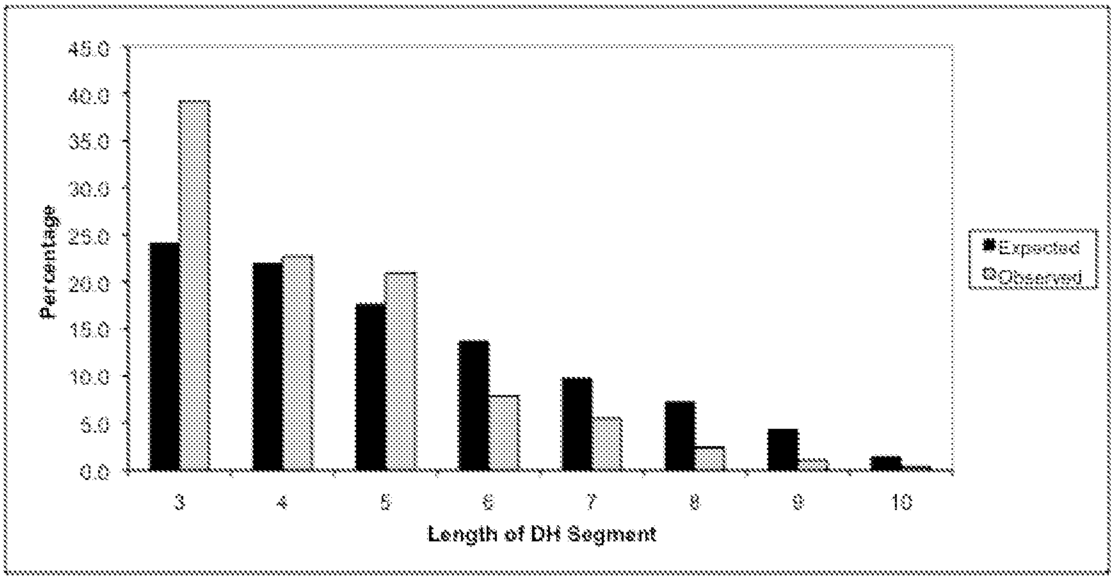

45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
3
4
5
6
7
8
9
10
Length of DH Segment
Expected
Observed

Figure 17

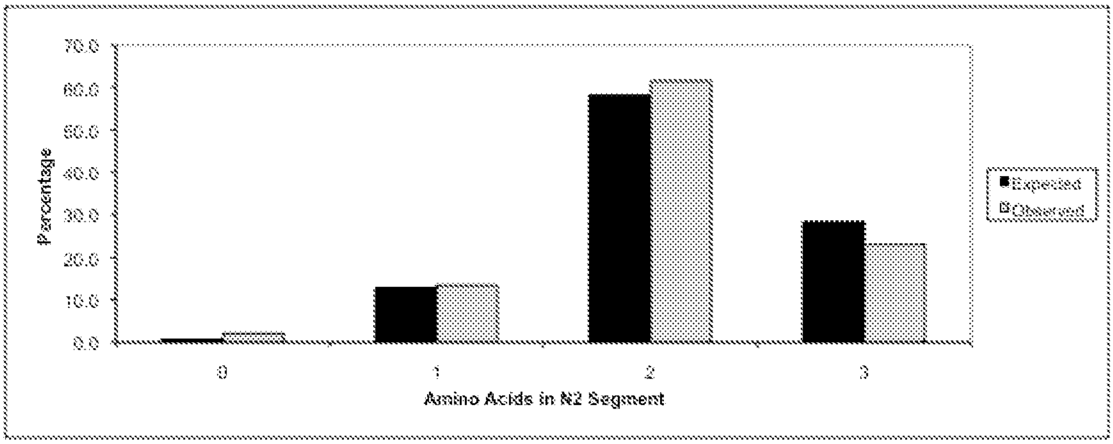

70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Percentage
0
1
2
3
Amino Acids in N2 Segment
■ Expected
▨ Observed

Figure 18

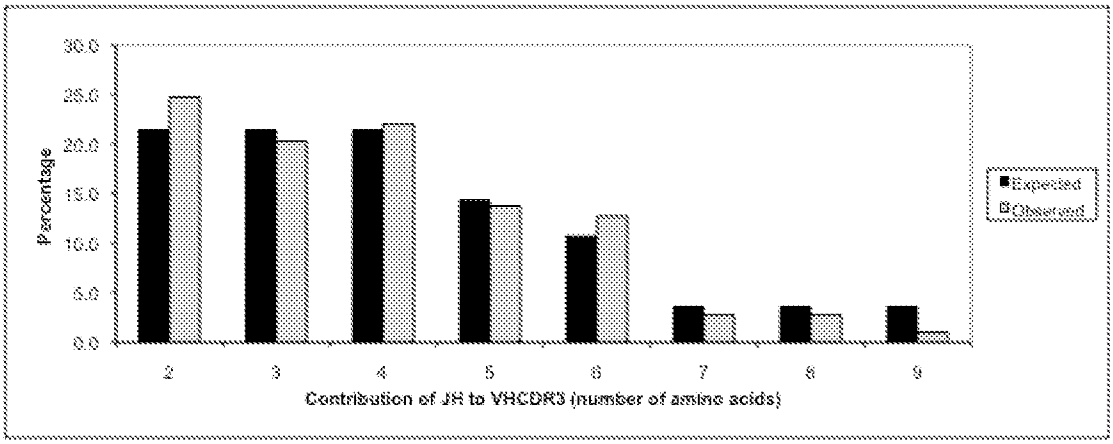

30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
2
3
4
5
6
7
8
9
Contribution of JH to VHCDR3 (number of amino acids)
Expected
Observed

Figure 19

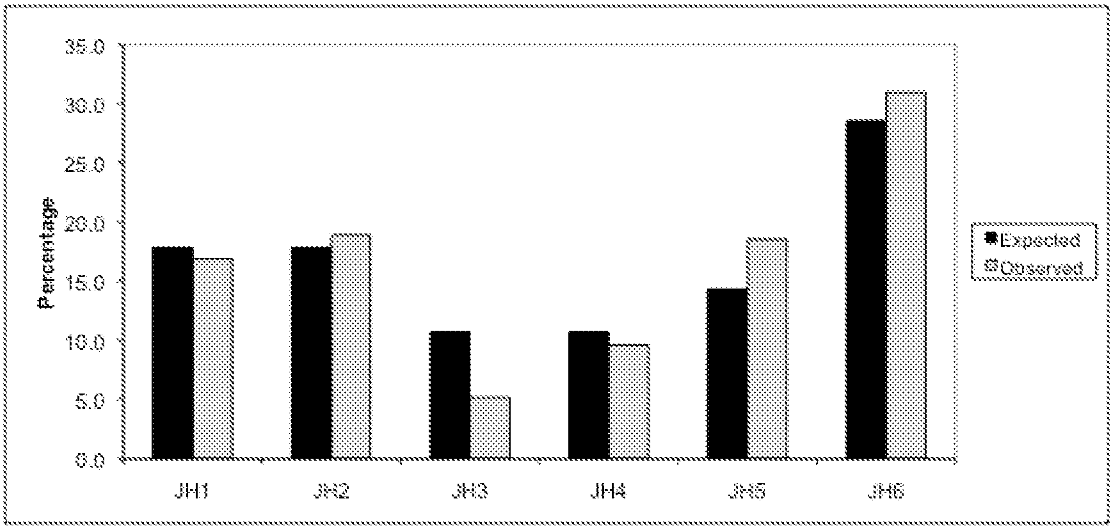

35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
JH1
JH2
JH3
JH4
JH5
JH6
Expected
Observed

Figure 20

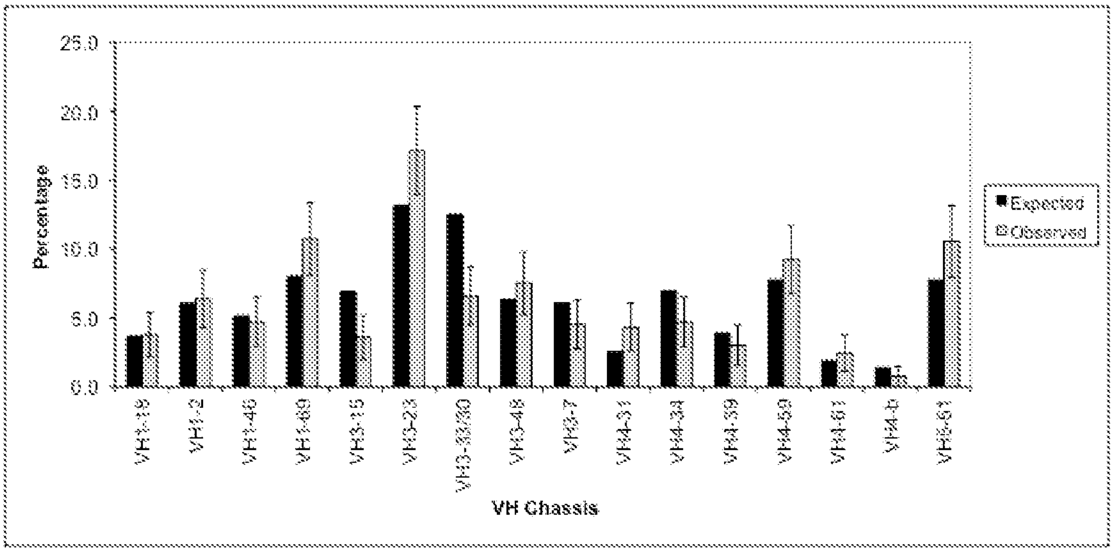

25.0
20.0
15.0
10.0
5.0
0.0
Percentage
Expected
Observed
VH1-18
VH1-2
VH1-46
VH1-69
VH3-15
VH3-23
VH3-33/30
VH3-48
VH3-7
VH4-31
VH4-34
VH4-39
VH4-59
VH4-61
VH4-b
VH5-51
VH Chassis

Figure 21

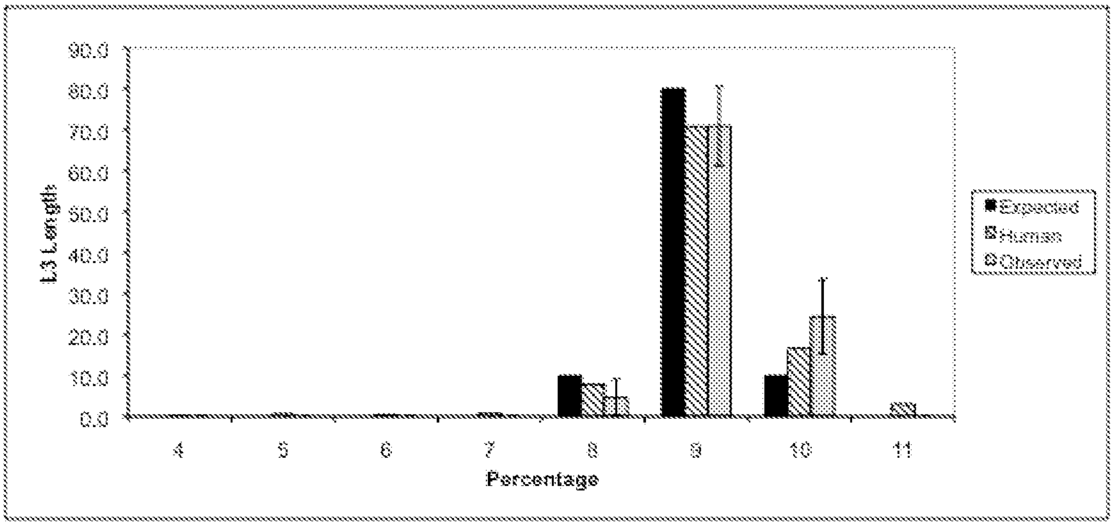

90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
L3 Length
4
5
6
7
8
9
10
11
Percentage
Expected
Human
Observed

Figure 22

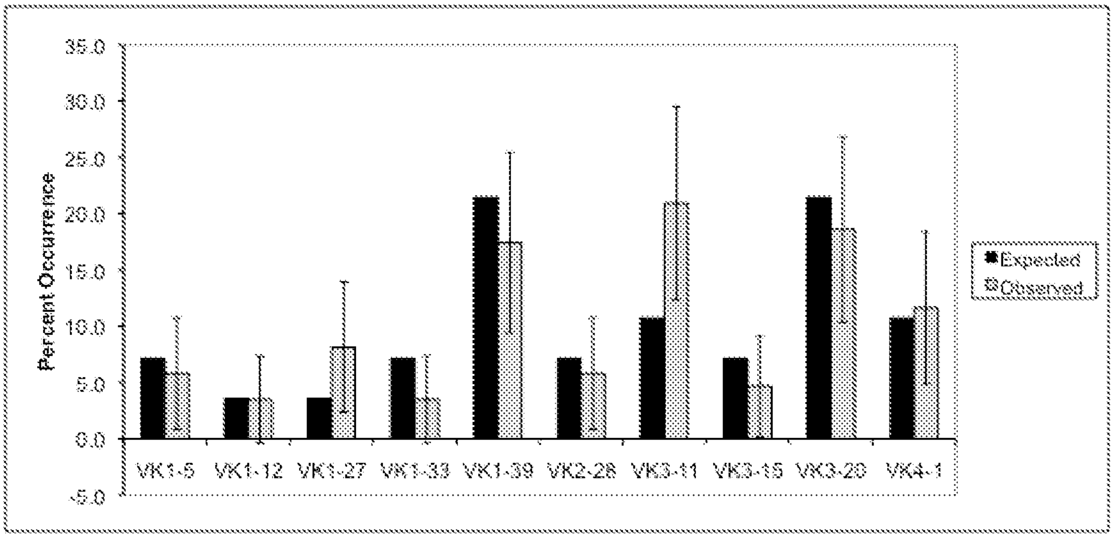

Percent Occurrence
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
-5.0
VK1-5
VK1-12
VK1-27
VK1-33
VK1-39
VK2-28
VK3-11
VK3-15
VK3-20
VK4-1
■ Expected
▨ Observed

Figure 23

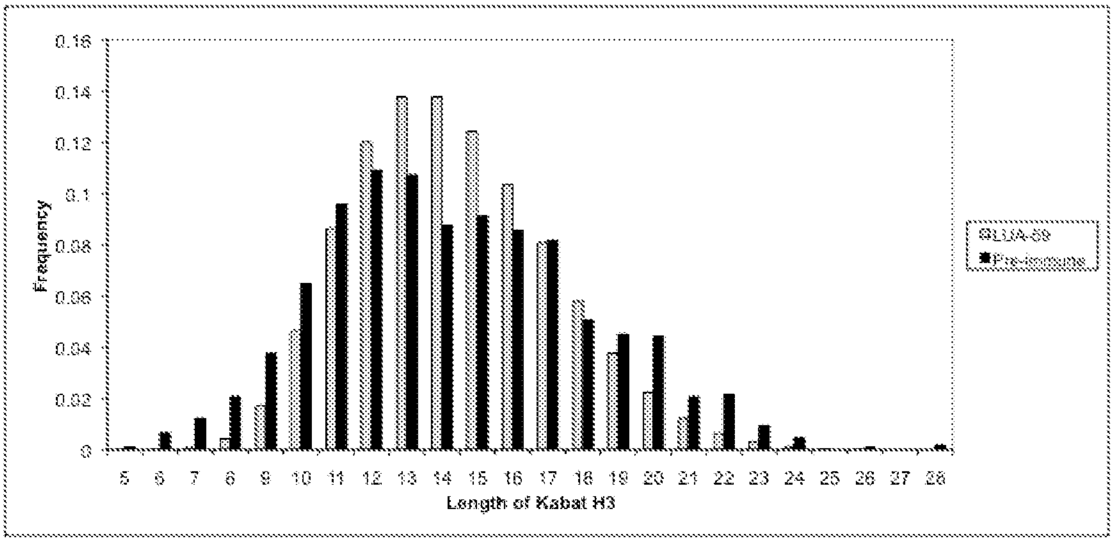
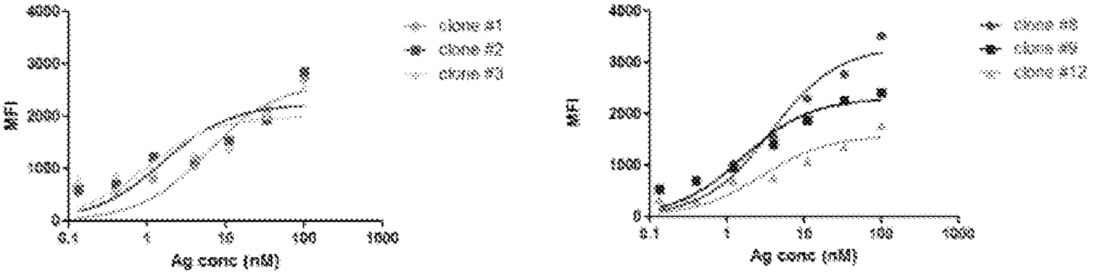

0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Frequency
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Length of Kabat H3
▨LUA-59
■Pre-immune

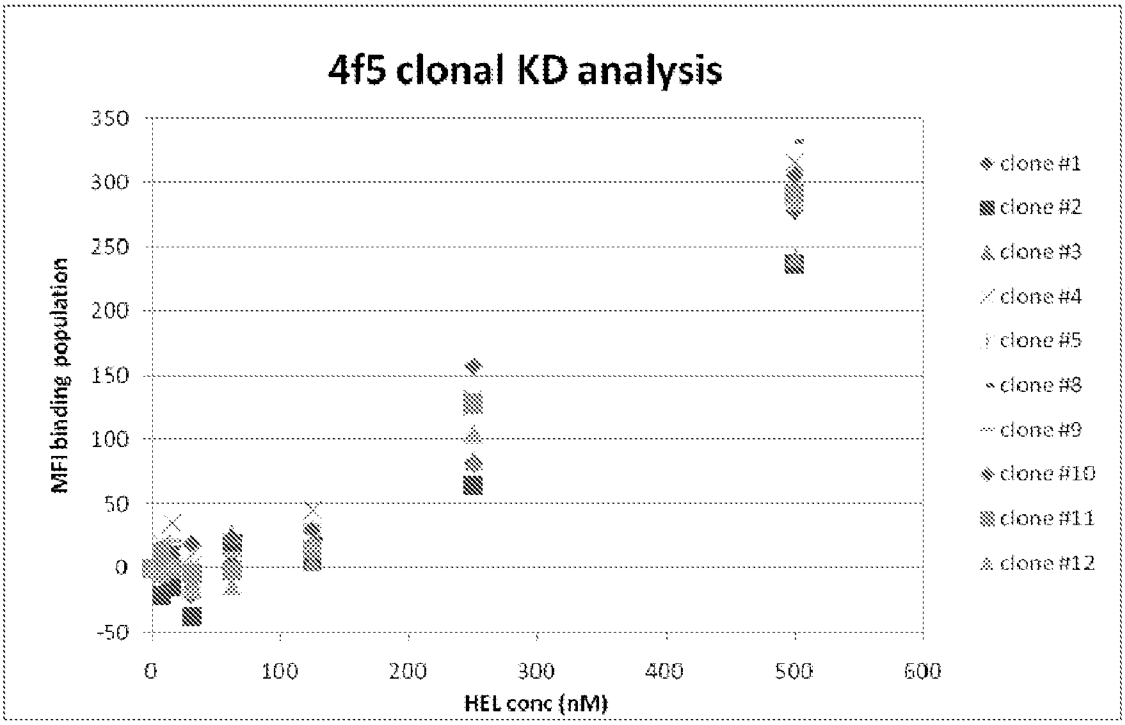
Figure 24



	clone #1	clone #2	clone #3
Kd	5.009	1.534	0.9139

	clone #8	clone #9	clone #12
Kd	3.823	1.576	2.752

Figure 25


4f5 clonal KD analysis
MFI binding population
350
300
250
200
150
100
50
0
-50
0
100
200
300
400
500
600
HEL conc (nM)
clone #1
clone #2
clone #3
clone #4
clone #5
clone #8
clone #9
clone #10
clone #11
clone #12