ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Rapid Construction and Characterization of Synthetic Antibody Libraries Without DNA Amplification

Xin Ge,[1,2] Yariv Mazor,[1,2] Scott P. Hunicke-Smith,[2] Andrew D. Ellington,[2,3] George Georgiou[1,2,4,5]

[1]Department of Chemical Engineering, University of Texas at Austin, Austin, Texas 78712; telephone: +1-512-471-6975; fax: +1-512-471-7963; e-mail: gg@che.utexas.edu

[2]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas

[3]Department of Chemistry and Biochemistry, Microbiology, University of Texas at Austin, Austin, Texas

[4]Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas

[5]Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas

Received 12 January 2010; revision received 10 February 2010; accepted 16 February 2010

Published online 2 March 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bit.22712

ABSTRACT: We report on a simple method to rapidly generate very large libraries of genes encoding mutant proteins without the use of DNA amplification, and the application of this methodology in the construction of synthetic immunoglobulin variable heavy (V_H) and light (V_κ) libraries. Four high quality, chemically synthesized polynucleotides (90–140 bases) were annealed and extended using T4 DNA polymerase. Following electroporation, $>10^9$ transformants could be synthesized within 1 day. Fusion to β-lactamase and selection on ampicillin resulted in 3.7×10^8 V_H and 6.9×10^8 V_κ clones highly enriched for full-length, in-frame genes. High-throughput 454 DNA sequencing of >250,000 V_H and V_κ genes from the pre- and post-selection libraries revealed that, in addition to the expected reduction in reading-frame shifts and stop codons, selection for functional expression also resulted in a statistical decrease in the cysteine content. Apart from these differences, there was a good agreement between the expected and actual diversity, indicating that neither oligonucleotide synthesis nor biological constrains due to protein synthesis of V_H/V_κ-β-lactamase fusions introduce biases in the amino acid composition of the randomized regions. This methodology can be employed for the rapid construction of highly diverse libraries with the near elimination of PCR errors in invariant regions.

Biotechnol. Bioeng. 2010;106: 347–357.

KEYWORDS: combinatorial DNA library; synthetic antibody; high-throughput sequencing; gene assembly; T4 DNA polymerase

Y. Mazor's present address is MedImmune, Gaithersburg, MD.
Correspondence to: G. Georgiou
Contract grant sponsor: Clayton Foundation

Introduction

The isolation of novel proteins from large combinatorial libraries has been greatly facilitated by the development of methods for library construction that rely on gene assembly from degenerate oligonucleotides and PCR amplification (Hoogenboom, 2005; Marks et al., 1991; Orlandi et al., 1989; Sidhu and Fellouse, 2006). Library quality and diversity are of particular importance in the isolation of antigen-specific antibodies from synthetic combinatorial libraries. Synthetic antibody libraries are constructed by introducing diversity in the complementarity determining regions (CDRs) of variable domains so that all or a subset of the 20 natural amino acids are represented at various positions known to be critical for antigen recognition (Cobaugh et al., 2008; Fellouse et al., 2007; Hoet et al., 2005; Knappik et al., 2000; Lee et al., 2004; Rajpal et al., 2005; Sidhu et al., 2004; Silacci et al., 2005). Typically, the randomized CDR regions are grafted either onto a single, or at most a small set, of constant framework regions (FRs) (Rothe et al., 2008). Because multiple sites in the gene have to be subjected to mutagenesis, the construction of synthetic antibody libraries relies on methods that capitalize on gene amplification, for example, total gene synthesis using degenerate primers or overlap extension PCR (Hayashi et al., 1994). However, synthesis errors and misincorporation of nucleotides during PCR amplification can often introduce unanticipated mutations (Cline et al., 1996; Pienaar et al., 2006); for example, in FRs of antibody variable domains (Clark, 2000). Such mutations are of concern for therapeutic applications because of the potential for eliciting adverse immune responses. Moreover, it can be time consuming and

technically difficult to obtain large amounts of DNA for the construction of large libraries ($>10^9$ transformants) by PCR-based methods.

We report the construction and characterization of highly diverse antibody libraries without PCR amplification by overlap extension of long oligonucleotides encoding diversified CDRs and FR gene fragments with the very high fidelity T4 DNA polymerase. A large amount of full-length library DNA is generated within hours allowing subsequent vector ligation and transformation steps more convenient and efficient. Libraries were subjected to in-frame selection via fusion to β-lactamase and the diversity of the clones pre- and post-selection was analyzed by high-throughput DNA sequencing using the Roche 454 technology. We find that expression of the V_H or V_κ genes did not substantially alter the amino acid distribution in the CDRs, which corresponded closely to the designed codon diversity. Notable exceptions were the expected depletion of stop codons following in-frame selection and also a statistically significant reduction in the Cys content in CDR3 of the V_κ but not of the V_H gene. The V_H or V_κ gene pools are suitable for construction of scFv, F_{AB}, or IgG (Mazor et al., 2008) synthetic libraries for antibody isolation by combinatorial screening methods.

Materials and Methods

V_H and V_κ Gene Assembly

The occurrence frequency of amino acid in natural human antibodies (germline V_HIII and V_κIII) were statistically analyzed using the KabatMan antibody database (Martin, 1996). Amino acid frequency analysis was performed for every position of CDR-L1, L2, H1, H2, and position L90, L92, L93 in CDR-L3 and H94 in CDR-H3, and the results are listed in Table I. An NNS codon was introduced to position L91, L94, L96 of CDR-L3, and H95-100c of CDR-H3 to encode all 20 amino acids. Codon usage of the FRs was optimized for *E. coli* expression (Puigbo et al., 2007). Four polynucleotide populations were designed to encode parts of FR1, FR1 + CDR1 + FR2, FR2 + CDR2 + FR3, FR3 + CDR3 + FR4, respectively, with overlapping regions at FR1, FR2, and FR3. The length and location of these overlap segments were carefully chosen to allow annealing at the same temperature (~58°C).

Overall, 12 polynucleotides with lengths between 90 and 140 bases (5 for V_κ and 7 for V_H) were synthesized (with 5′ phosphorylation for oligonucleotides VH2, VH3, V_κ2a/b, and V_κ3), and PAGE purified with typical yields of ~2 nmol (IDT, Coralville, IA). Equimolar amounts of oligonucleotides (25 pmol for VH4a/b/c/d, and 50 pmol for V_κ2a/b) were annealed in 50 mM NaCl, 10 mM Tris–HCl by gradually cooling the primer mixtures from 95 to 25°C. Gap filling was performed using T4 DNA polymerase and T4 DNA ligase in the presence of 500 μM dNTPs for 60 min at 37°C followed by heat inactivation at 75°C for 20 min in the presence of 10 mM EDTA. The assembled V_H/VK genes were gel purified using DNA recovery kit (Zymo Research, Orange, CA).

In-Frame Selection by Expression of V_H/V_κ β-Lactamase Fusions

Two β-lactamase fusion vectors, pVH-bla and pVL-bla, were constructed to facilitate the selection of in-frame library members able to express full-length variable domains. The lac promoter and pelB leader peptide were used for expression and secretion, respectively, of V_H-bla and VL-bla fusion proteins into periplasm. Briefly, the 3,748 bp fragment (*Nde*I/*Hin*dIII ended) of pMoPac1 (Hayhurst et al., 2003) and the 427 bp fragment (*Nde*I/*Hin*dIII ended) of pMAZ360-IgG (Mazor et al., 2007) were ligated to give pVH. The *bla* gene, encoding β-lactamase, was amplified by PCR with the primers 5′-gcataagcttcggccacccagaaacgctggtgaaag-3′ and 5′-cgcagtctg-gatccgtttaagggcaccaataactgc-3′ and the template pMAZ360-IgG. The PCR product was then digested with *Hin*dIII and *Bam*HI, and inserted into the same sites on pVH, yielding pVH-bla. For the construction of pVL-bla, the 3,921 bp fragment (*Nde*I ended) of pMoPac1 was self-ligated to give pMoPac100. The *Nco*I site on pMoPac100 was removed by QuickChange (Stratagene, La Jolla, CA) site directed mutagenesis using primers 5′-cttcgcccccgttttcacaatgggcaaatattatac-3′ and 5′-gtataatatttgcccattgtgaaaacgggggcgaag-3′, and resulted in pMoPac99. The 3,637 bp fragment (*Kpn*I/*Not*I ended) of pMoPac99 was then ligated with the 608 bp fragment (*Kpn*I/*Not*I ended) of pMAZ360-IgG to give pVL. The *bla* gene was PCR amplified with primers 5′-catagcggccgctcacccagaaccgctgg-3′ and 5′-cgcagtctggatccgtttaagggcaccaataactgc-3′, and then digested with *Not*I/*Bam*HI. This fragment was inserted into the same sites on pVL, and cultured on Cam+/Amp+ plate to give the selection vector pVL-bla. The assembled V_H/V_κ genes and also the selection vectors pVH-bla and pVL-bla were subjected to double restriction digestion by *Nhe*I/*Hin*dIII and *Nco*I/*Not*I respectively. DNA was gel purified, desalted and finally, 1 pmol of vector and 2 pmol of assembled V_H/V_κ DNA were ligated, desalted, and introduced to *E. coli* by electroporation (Mazor et al., 2008).

E. coli stain Jude-1, DH10B (Invitrogen, Carlsbad, CA) harboring the "F" factor derived from XL1-blue (Stratagene), was used for library construction. Cell growth and protein expression conditions were optimized to balance efficient in-frame selection and the ability to generate large numbers of transformants. Briefly, different concentration of inducer (0.2–1 mM isopropyl β-D-1-thiogalactopyranoside, IPTG) and antibiotics (25–200 μg/mL ampicillin) supplemented in 2 × YT agar media were tested, the number of transformations were counted following serial dilution and ~16 clones from each batch were randomly picked and sequenced. Using the optimized conditions (0.5 mM IPTG and 50 μg/mL ampicillin), transformation of 300 OD electrocompetent cells with 3 μg ligation typically produced a yield of $\sim 10^9$ transformants.

Table I. Prevalence of aa at different positions within the light and heavy chain CDRs of human antibodies (V_κIII, V_HIII), and sequence diversification scheme used for library construction.

CDR (length)	Position	Prevalence of different amino acids at different positions within human antibodies (%)	Diversity design: Codon[a]	Residues encoded	Coverage (%)	Oligo
L1 (11)	L29	V (60); I (37)	RTT	VI	97	VL2a
	L30	S (76); G (13); N (5); R (2); T (1)	RGC	SG	89	
	L31	S (74); N (10); T (7); Y (1); G (1); R (1); K (1); D (1)	AVC	SNT	90	
	L32	Y (59); N (32); S (5); F (3)	WAT	YN	90	
L1 (12)	L28	V (87); I (8)	RTT	VI	94	VL2b
	L29	S (87); T (3); N (2); R (2); D (2)	AVC	STN	93	
	L30	S (82); N (9); R (6); G (2); A (2)	ARC	SN	91	
	L31	S (69); N (14); T (9); G (2); A (2); R (2)	AVC	SNT	92	
L2	L50	G (45); D (39); A (12); S (3); Y (2); E (1)	GVC	GDA	90	VL3
	L51	A (88); T (8); V (1)	RCC	AT	96	
	L53	S (37); N (37); T (18); K (4); R (1)	AVH	SNTKR[b]	96	
L3[c]	L90	Q (94); H (6)	CAD	QH[d]	99	VL4
	L91	Xaa	NNS	Xaa	—	
	L92	G (49); S (30); N (9); D (3); E (3); R (2)	RRC	GDND	91	
	L93	S (37); N (31); T (10); A (5); G (2); D (2)	RVC	SNTAGD	86	
	L94	Xaa	NNS	Xaa	—	
	L96	Xaa	NNS	Xaa	—	
H1	H31	S (46); N (16); D (13); T (9); G (8); R (3); A (1); K (1)	RVC	SNDTGA	94	VH2
	H32	Y (74); A (6); S (5); F (4); H (3); N (3); V (2); D (1)	KMC	YASD	86	
	H33	A (31); W (19); Y (16); G (15); S (6); E (4); D (3) T (2)	KSG	AWGS	70	
	H35	S (41); H (31); N (16); T (6); Y (1)	HMT	SHNTY P	96	
H2	H50	V (19); A (16); Y (13); N (11); G (9); S (8); L (5); W (4); T (3); R (3); I (2); F (2); K (1)	DHC	VAYNSTIF D	74	VH3
	H52	S (60); K (11); N (9); W (6); T (3); R (3); Y (2)	ARH	SKNR[e]	83	
	H52a	G (21); Y (20); S (18); Q (9); P (5); N (4); W (4); F (3); A (3); D (2); T (2); E (2); R (2); H (1); K (1)	NMT	YSPNADTH	56	
	H53	S (34); D (39); N (10); G (8); R (3); T (1); Y (1); H (1)	RRT	SDNG	91	
	H54	G (75); S (16)	RGC	GS	90	
	H55	S (38); G (32); T (8); D (7); N (3); A (2); E (1)	RRC	SGDN	81	
	H56	S (26); N (19); T (15); E (10); Y (7); D (4); G (4); K (2); R (2); A (1); Q (1)	WMC	SNTY	67	
	H57	T (38); K (34); I (19); R (2)	AHA	TKI	90	
	H58	Y (68); F (5); S (4); H (4); T (1); R (1)	TWT	YF	73	
H3 (9–12)	H94	R (56); K (22); T (9)	ARA	RK	78	VH4a-d
H3 (12)	H95-100c	$(Xaa)_9$	$(NNS)_9$	$(Xaa)_9$	—	VH4d
H3 (11)	H95-100b	$(Xaa)_8$	$(NNS)_8$	$(Xaa)_8$	—	VH4c
H3 (10)	H95-100a	$(Xaa)_7$	$(NNS)_7$	$(Xaa)_7$	—	VH4b
H3 (9)	H95-100	$(Xaa)_6$	$(NNS)_6$	$(Xaa)_6$	—	VH4a

[a]IUB code: B = C/G/T, D = A/G/T, H = A/C/T, K = G/T, M = A/C, N = A/C/G/T, R = A/G, S = G/C, V = A/C/G, W = A/T, Y = C/T.
[b]The designed molar ratio of Thr/Ser/Asn/Lys/Arg = 3:2:2:1:1.
[c]L95 = P and L97 = T.
[d]The designed molar ratio of Gln/His = 2:1.
[e]The designed molar ratio of Ser/Lys/Asn/Arg = 2:2:1:1.

High-Throughput Sequencing of V_H/V_κ Genes

100 OD (>20 times the number of transformants) of *E. coli* cells carrying selected V_H/V_κ gene libraries, were inoculated into 500 mL LB supplemented with 30 μg/mL chloramphenicol, 0.6% glucose, and cultured at 30°C for 6 h. Plasmid DNA was extracted using the Maxiprep Kit (Qiagen, Valencia, CA), digested with *Nhe*I and *Hin*dIII for V_H, and *Nco*I and *Not*I for V_κ, gel purified and concentrated to >100 ng/μL. Four micrograms V_H/V_κ DNA fragments from pVH-bla and pVL-bla (denoted as post-selection), and 2 μg of assembled V_H/V_κ (denoted as pre-selection), were submitted for high-throughput sequencing (SeqWright, Houston, TX). The Roche 454 Titanium sequencer was run with each sample occupying one eighth of one picotiter plate, and standard primer and quality trimming were performed. Analysis was performed using Perl scripts in a Unix environment. Data were grouped into V_H-like and V_κ-like fragment pools by similarity search of the FRs. Local sequences adjacent to CDRs were then aligned to identify all of the six CDRs of both V_H and V_κ.

Results

Design of Antibody Complementarity Determining Regions (CDRs) Amino Acid Diversification

The human antibody germline variable domains, DP47 (V_HIII) and DPK22 (V_κIII) were used as the scaffold to

construct synthetic antibody libraries because (1) they are highly prevalent (i.e., 12% and 29% of known human antibodies, respectively) (Knappik et al., 2000); and (2) they are well-expressed in bacteria due to their thermodynamic stability (Ewert et al., 2003). The synthetic library was designed to randomize all six CDRs on both V_H and V_κ.

The CDR-H3 was designed to have a length between 9 and 12 amino acids, which matched the mean distribution of CDR-H3 length in the KabatMan and IMGT databases (Zemlin et al., 2003). CDR-H3 regions with different lengths were encoded using four individual polynucleotides (VH4a/b/c/d) having 6, 7, 8, and 9 NNS codons respectively at H95-H100c, followed by the highly conserved sequence Phe-Asp-Tyr (Silacci et al., 2005). Within V_κ, two randomized CDR-L1 regions with different lengths (four or five amino acids) were generated, positions L91, L94, and L96 were designed to be encoded by NNS, and positions L95 and L97 were fixed to Pro and Thr, which are highly conserved in V_κ (Martin, 1996; Silacci et al., 2005).

Amino acid frequencies at other CDR positions were determined for human V_HIII and V_κIII antibodies in the KabatMan antibody database and served as the basis for the design of synthetic, degenerate codons. Codon diversification schemes that precluded stop or cysteine residues were used where possible (Lee et al., 2004; Liang et al., 2007). As shown in Table I, the degenerate codons that were eventually employed allowed high coverage (~92% on average) of the desired amino acids in CDR-L1, L2, and L3. For the majority of positions in V_H, the randomization scheme allowed greater than 80% coverage of the natural amino acid diversity found in the database. Restrictions on codon degeneracy reduced coverage at only two positions, H52a and H56 (56% and 67%, respectively). Diversity could potentially be further expanded or refined using oligonucleotides synthesized from trinucleotide building blocks (Knappik et al., 2000; Liang et al., 2007; Rothe et al., 2008) but at a significantly higher cost.

The complete diversification scheme for V_H and V_κ, including FRs and CDRs, is summarized in Figure 1, which also includes the degenerate codons used. The numbers of residues diversified in V_κ, CDR-1, 2, and 3 were 4, 3, and 6, respectively and 4, 9, and 7–10 for V_H, CDR 1, 2, and 3. The theoretical sequence diversity was calculated to be 6.9×10^8 for V_κ, and $>10^{19}$ for V_H.

Assembly of Antibody Variable Domain Genes (V_H and V_κ)

As demonstrated in Figure 2, the construction of libraries by annealing and gap filling is very simple and can be completed within a few hours. Four oligonucleotides of between 90 and 140 nt in length and codon-optimized for expression in *E. coli* were synthesized and gel purified. When these four oligos are annealed, they form hemiduplexes in which the FRs are base-paired while the randomized CDRs remain as single-stranded "gaps" to be filled in with T4 DNA polymerase. T4 DNA polymerase was used instead of other mesophilic polymerases (e.g., Klenow, ϕ29) because it lacks strand displacement activity. T4 DNA ligase was also added to the reaction mixture to seal the nicks remaining

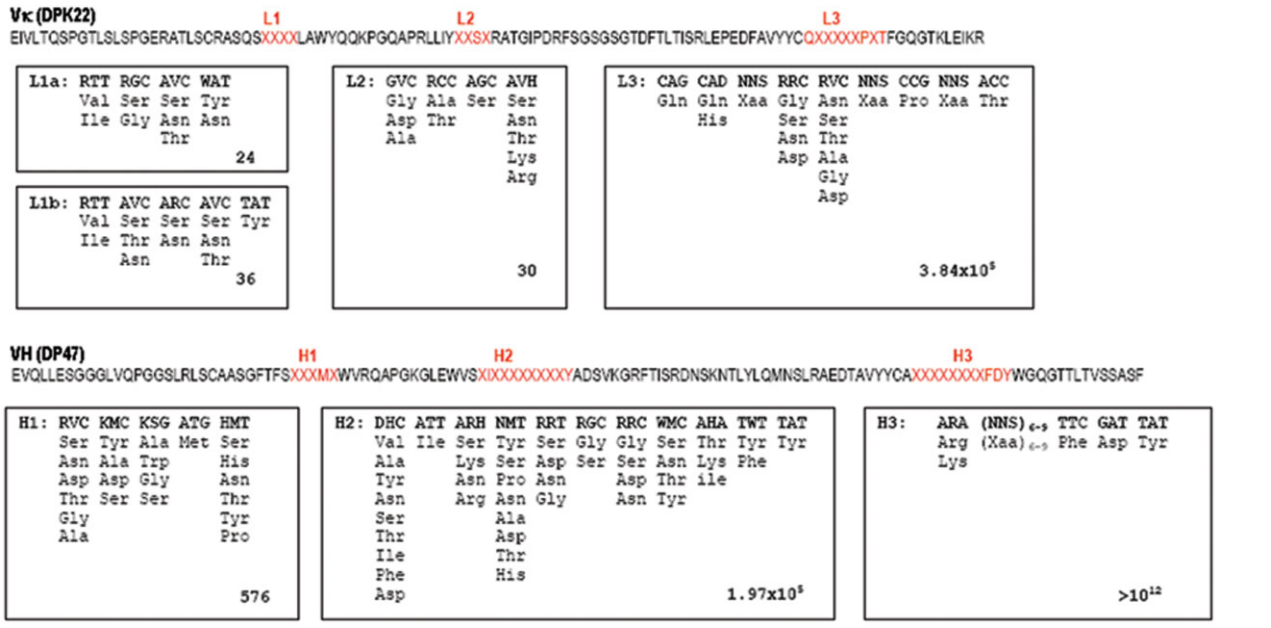

Figure 1. Amino acid and nucleotide sequences of V_H and V_κ. CDR regions subjected to diversification are shown in red and the corresponding IUB codon schemes are shown in the corresponding boxes. The theoretical sequence diversity of each diversified region is shown at the lower right side in each box. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

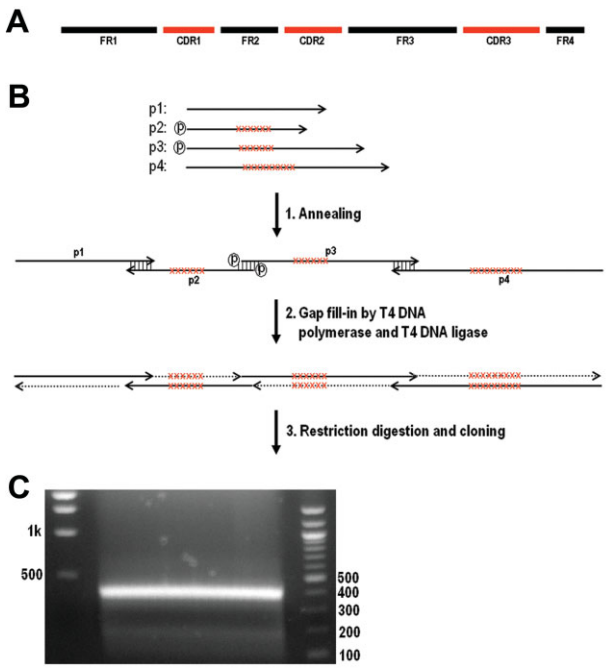

Figure 2. Construction of V_H and Vκ gene libraries. **A**: Structure of antibody variable domain gene. The red regions represent CDRs. **B**: Gene assembly by annealing and gap filling. Oligos of length between 90 and 140 nt were designed to contain complementary sequences within the FRs, while CDR1, CDR2, CDR3 were encoded by primer 2, primer 3, primer 4, respectively. Oligos 2 and 3 were 5′-phosphorylated. After annealing, the gaps were filled by treatment with T4 DNA polymerase and T4 DNA ligase simultaneously, generating double-stranded V_H/V_κ fragments ready for restriction digestion and cloning. **C**: DNA product of the gap-filling reaction for the V_H gene. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

after DNA polymerization. The assembled V_H and V_κ fragments were analyzed by gel electrophoresis and a single band of the appropriate size was observed (Fig. 2C). Reactions using 100 pmol of oligonucleotides typically yielded 15–18 μg of high purity, fully assembled, genes. The large amount of DNA obtained ($>10^{13}$ molecules) makes the subsequent ligation steps more conveniently, and eventually allows the construction of highly diverse libraries. The external oligonucleotides also contain a unique restriction site for cloning of the full segment of approximately 350 bp into the pVH-bla and pVL-bla in-frame selection vectors. Libraries with $>10^9$ transformants were routinely generated using V_H/V_κ genes assembled by this annealing and extension method.

In-Frame Selection of V_H and V_κ Libraries

Sequencing results from randomly picked colonies revealed that only ~60% (12/19) of the V_H clones carried full-length ORFs due to the presence of stop codons (3/19) within CDR-H3 or deletions (4/19) within FRs. The deletions are likely a consequence of inefficient coupling during oligonucleotide synthesis. To reduce the incidence of frame shifts and stop codons, the V gene libraries were subjected to in-frame selection by constructing C-terminal fusions to β-lactamase and selection for resistance to ampicillin (Lutz et al., 2002; Rothe et al., 2008; Seehaus et al., 1992). Briefly, the in-frame selection vectors, pVH-bla and pVL-bla, were constructed and their utility for in-frame selections was validated using known V_H clones (data not shown). Cell growth and protein expression conditions, such as the concentration of inducer (IPTG) and antibiotics (ampicillin), were optimized to balance efficient in-frame selection and the ability to generate large numbers of transformants (details in the Materials and Methods Section). Following in-frame selection the size of the resulting libraries was estimated by plating serial dilutions of transformants. A total of $3.7 \pm 0.3 \times 10^8$ transformants were obtained for the V_H library, and $1.7 \pm 0.2 \times 10^9$ for the V_κ library. The V_κ library therefore had >2 fold coverage of the theoretical diversity of 6.9×10^8 variants, while the most diverse V_H library covered only a fraction of its potential sequence space.

To evaluate the quality of the gap-filling reaction for gene assembly, 100 clones were randomly picked from V_H and V_κ libraries and the variable genes were sequenced by conventional Sanger sequencing in both directions. It was found that 100% of the V_H genes (54/54), and ~93% of the V_κ genes (43/46) were full-length with no mutations in the invariable regions. All three V_κ mutant sequences (not full-length) contained single nucleotide deletions; no transitions or transversions, the major types of mutations by T4 polymerase were found (Kunkel et al., 1984). Therefore, it is unlikely that the errors were introduced during the extension performed by T4 DNA polymerase. More stringent selection conditions could be used to further reduce the frequency of frame-shifted clones in the V_κ gene library.

Library Characterization by High-Throughput Sequencing

High-throughput (454) sequencing technology was exploited to characterize sequence coverage within the V_H and V_κ libraries. Two samples were prepared and analyzed: assembled V_H/V_κ genes immediately following annealing and gap filling (denoted as pre-selection), and gene fragments recovered from pVH-bla or pVL-bla transformants (denoted as post-selection). Quality trimmed data consisted of 210,000 and 96,653 sequences for pre-selection and post-selection samples, respectively. Read length distributions clearly showed two sequence clusters at ~340 and ~390 nt, corresponding to expected lengths of V_κ and V_H. Sequences were grouped into V_H-like and V_κ-like pools, and the six CDRs were identified (Fig. 3 for the entire programming flowchart of the data mining process). Statistical analyses (Table II) show there are an average of ~48,000 sequences for each CDR, with an even distribution of different designed lengths for the 6–9 NNS codons in CDR-H3 (25.0%, 29.3%, 25.1%, 20.7%).

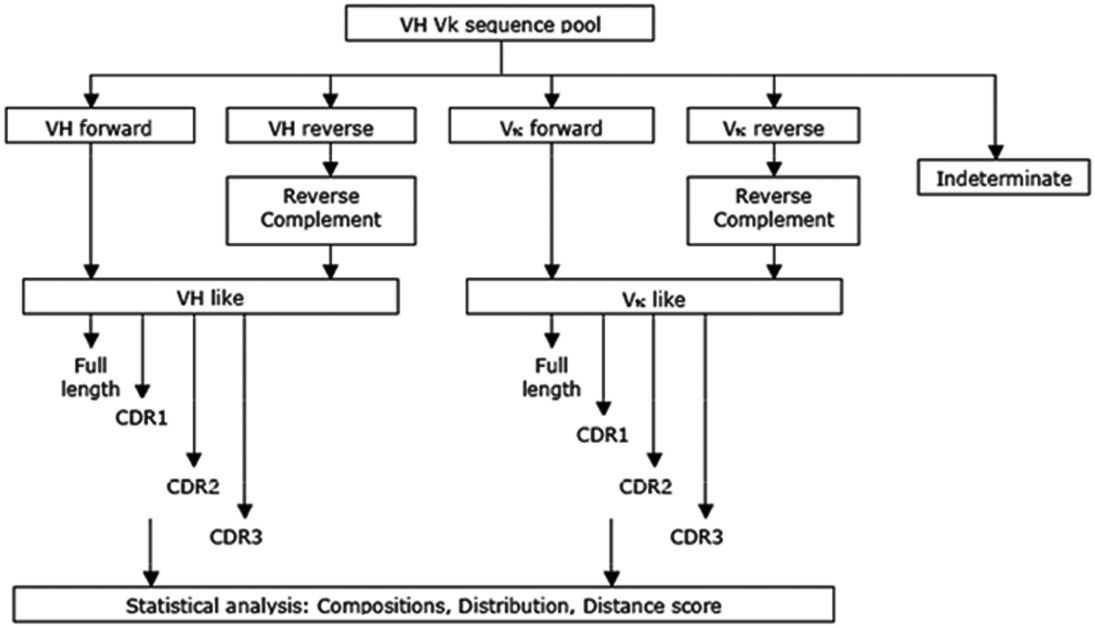

Figure 3. Programming flow chart. Data were grouped into V_H-like and V_κ-like fragment pools by homology similarity search of the framework regions. Local sequences adjacent to CDRs were then aligned to identify all of the six CDRs of both V_H and V_κ.

A preliminary statistical analyses reveals that, in V_H FR 3 (FR-H3, one of the longest FRs) the majority of mutations (>67%) were associated with homopolymers, especially at penta-adenine (AAAAA) segments. This agrees with literature that the most common errors generated during 454 sequencing occur at nucleotide repeats, due to difficulties in resolving the luminescence intensity (homopolymer-length errors), insufficient flushing between nucleotide extensions (base insertions near homopolymers), or insufficient nucleotide concentration during extension (incomplete extension) (Huse et al., 2007). In fact, the error rate of the 454 sequencer ($\sim 10^{-3}$) is roughly four magnitudes higher than the fidelity of T4 DNA polymerase (Huse et al., 2007; Kunkel et al., 1984). Thus, while the data are not of sufficient accuracy to deduce incorporation errors that may have been generated during the gap-filling reaction, it is suitable for the analysis of amino acid diversity coverage.

Table II. Numbers of identified V_H/V_κ genes and CDRs.

	Pre-selection	Post-selection
V_κ	112,829	53,517
CDR-L1		
11aa	44,518	23,925
12aa	50,878	18,889
CDR-L2	96,120	43,340
CDR-L3	80,537	38,356
V_H	54,902	36,090
CDR-H1	43,034	27,155
CDR-H2	36,609	26,439
CDR-H3		
9aa	8,069	6,072
10aa	8,122	7,122
11aa	8,517	6,098
12aa	7,535	5,024

Amino acid diversity coverage was evaluated first by analyzing the amino acid frequency at each individual residue within the CDRs. This analysis revealed that there was no significant deviation between the programmed and experimentally determined diversity for CDR-L1, L2, H1, and H2 (Fig. 4). In addition, for the NNS codons within CDR-L3 and H3, all 20 amino acids were found at the expected proportions (Fig. 5). As expected, in-frame selection resulted in significant stop codon depletion. For instance, in the six NNS sub-library of CDR-H3 the stop codon content at each position was about 3.5% on average before in-frame selection (consistent with a theoretical probability of 1/32, for the NNS randomization scheme), while this number dropped to ~0.1% following in-frame selection. Overall, before in-frame selection approximately 18–28% CDR-H3 sequences were found to contain stop codons (Table IIIA), which well agrees with the theoretical probability (17–25%, depending on the number of random positions). After in-frame selection, the frequency of stop codons decreased to ~0.6% on average for $(NNS)_{6-9}$, a more than 30-fold reduction. A somewhat smaller reduction in the frequency of stop codons was observed in the V_κ library (Fig. 5B and Table IIIA).

It has been demonstrated that the most common error in chemical oligonucleotide synthesis is base deletion leading to frame shifts (Hecker and Rill, 1998). Further analyses were preformed to evaluate the efficiency removal of

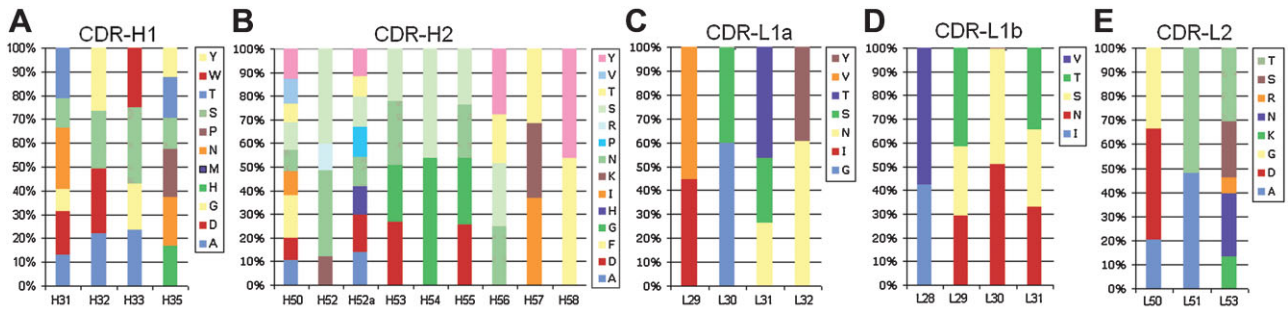

Figure 4. Amino acid composition of CDR-H1, L1, and L2 following in-frame selection. The sequence sample sized were 27155, 26439, 23925, 18889, 43340 for CDR-H1, H2, L1a (11aa), L1b (12aa), and L2, respectively. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

frame-shifted sequences by β-lactamase-based selection (Table IIIB). As an example, in CDR-H2 (which is one of the longest designed CDR sequences) full-length sequences (30 bases) and deletion variants (29, 28 bases) were counted in both pre-selection and post-selection samples. 6.3% of CDR-H2 sequences were frame shifted in the pre-selection sample, compared to 0.9% in the post-selection sample.

It was found that among three NNS positions in CDR-L3, the Cys content decreased from $2.93 \pm 0.17\%$ pre-selection, to $0.56 \pm 0.05\%$ in the post-selection sequence set (Fig. 5B). A less dramatic decrease in Cys content was also observed in CDR-H3, from $2.65 \pm 0.02\%$ to $1.80 \pm 0.12\%$. These results are consistent with previous studies that showed that Cys is rarely presented in antibody V genes (Knappik et al., 2000; Lee et al., 2004; Liang et al., 2007). Selection for expression of the β-lactamase fusion is expected to eliminate sequences that might be toxic to the cell, that block secretion, or that result in aggregation, and the presence of unpaired Cys residues could have been particularly problematic due to its propensity to form aberrant disulfide bonds. Interestingly, we noted a statistically significant increase in the Pro content in NNS encoded positions in CDR-H3. No statistically significant differences were observed in the relative abundance of all the other amino acids within CDR-L3 and H3.

Nearest neighbor analysis of dipeptide sequences was performed for CDR-H3 segments encoded by NNS. Occurrence frequencies of all 400 possible dipeptide sequences (20×20), and frequency ratio (R) between the post- and pre-selection data sets (24,316 and 32,243 sequences) were calculated (Table IV). Approximately 70% of the dipeptide sequences exhibited no significant difference between the pre- and post-selection samples (ratio of pre–post-selection frequency 0.8–1.2). Occupancy frequencies of dipeptides containing Cys, Phe, Ile, Met, Val, or Trp were found to decrease after selection whereas the frequency of dipeptides with Asp, Glu, Gly, His, Asn, Pro, or Gln was increased. Trp-Cys and Trp-Trp dipeptides showed

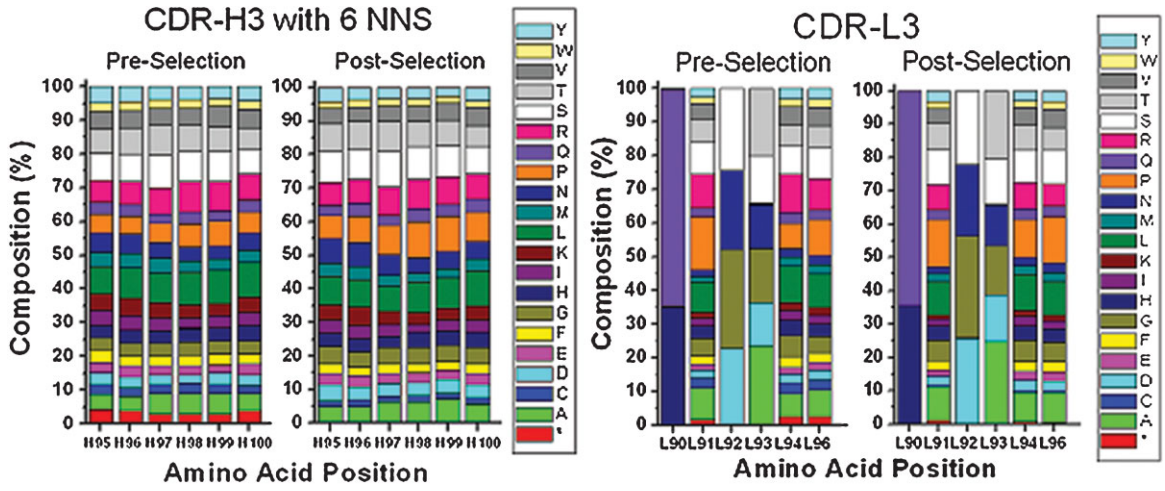

Figure 5. Statistical analysis of amino acid composition of CDR3s before and after in-frame selection. **A**: Amino acid composition of CDR-H3 with 6 aa randomized by an NNS scheme. Sequence sample sizes were 8069 and 6072 for pre- and post-selection, respectively. **B**: Amino acid composition of CDR-L3. Sequence sample sizes were 80,537 and 38,356 for pre–post-selection, respectively. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

Table III. Improvement of library quality by β-lactamase selection.

CDRs	Predicted % of seq. with stop codons[a]	Pre-selection: Number of seq. with stop codons	Pre-selection: Total number of seq.	Pre-selection: % of seq. with stop codons	Post-selection: Number of seq. with stop codons	Post-selection: Total number of seq.	Post-selection: % of seq. with stop codons	Improvement
(A) Removal of stop codon in CDR3s								
V_H-6NNS	17.3%	1,526	8,069	18.9%	35	6,072	0.58%	32.6
V_H-7NNS	19.9%	1,469	8,122	18.0%	32	7,122	0.45%	40.0
V_H-8NNS	22.4%	1,719	8,517	20.2%	42	6,098	0.69%	29.3
V_H-9NNS	24.9%	2,074	7,535	27.5%	39	5,024	0.78%	35.3
V_κ	9.1%	5,355	80,537	6.65%	803	38,356	2.09%	3.18

CDRs	Pre-selection: Number of seq. with −2 frame	Pre-selection: Number of seq. with −1 frame	Pre-selection: Number of seq. with 0 frame	Pre-selection: % of frame-shifted seq.	Post-selection: Number of seq. with −2 frame	Post-selection: Number of seq. with −1 frame	Post-selection: Number of seq. with 0 frame	Post-selection: % of frame-shifted seq.	Improvement
(B) Removal of frame-shifted CDR sequences									
L2	91	978	93,520	1.13	11	192	42,636	0.47	2.38
L3	710	9,044	76,972	11.2	170	1,317	37,208	3.84	2.92
H1	42	706	41,154	1.79	45	225	26,382	1.01	1.76
H2	132	2,284	35,635	6.35	17	224	25,791	0.93	6.86
H3	1,807	4,730	29,278	18.3	244	1,118	22,192	5.70	3.20

[a]The theoretical possibility to have stop codons in (NNS)n is calculated based on equation, $p(n) = 1 - (31/32)^n$.

the greatest depletion following screening whereas, Pro-Pro (ratio = 2.08) and Pro-His (ratio = 1.88) showed the greatest increase. Overall, dipeptides with decreased frequency contained hydrophobic residues, while dipeptide containing hydrophilic residues were enriched post-selection, most likely due to improved solubility and expression of the variable domains.

Distance score profiles were used to estimate the combinatorial distribution of amino acids in the CDRs. The Hamming distance between CDR sequence pairs was calculated: if two sequences have the same amino acid at position i, then $d_i = 0$; otherwise $d_i = 1$. The distance score is defined as, $D = \Sigma d_i$. For example for a five residue sequence, $D = 0$ indicates exactly same pentapeptide; and $D = 5$ indicates that all five positions are occupied by different amino acids. Distance score profiles were constructed for CDRs, and these experimental profiles were compared with theoretical profiles obtained by averaging 10 simulation runs of similar but randomly generated CDR sequences. The distance score profiles of CDR-L3, H2, and H3 were analyzed and some results are displayed in Figure 6. The experimentally determined distance scores are essentially identical to the simulation results with one exception: in the CDR-H3 library with nine positions randomized by NNS there were 40 pairs of identical amino acid sequences ($D = 0$ score) out of a sample size of 5,024 samples (Fig. 6A). Further characterization of these duplicates revealed that: (1) the entire nt DNA sequences in these 40 pairs of genes

Table IV. Frequency ratio (post-selection/pre-selection) of bi-amino acid in CDR-H3 encoded by $(NNS)_7$ scheme.

0.6 0.8 1.2 1.4 1.6

1st aa / 2nd aa	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.16	0.64	1.16	0.82	0.57	1.11	0.89	1.08	1.05	0.94	0.96	1.14	1.44	1.01	0.98	1.12	1.04	0.94	0.77	1.07
C	0.80	0.72	0.64	0.86	0.92	0.54	0.71	0.50	0.76	0.54	0.86	0.70	0.89	0.86	0.70	0.68	0.57	0.67	0.39	0.67
D	1.19	0.74	1.42	1.47	1.00	1.26	1.42	1.14	1.17	0.97	0.79	1.04	1.50	1.28	0.90	1.21	1.41	1.11	0.81	1.16
E	1.50	0.85	1.25	0.76	1.11	1.10	0.95	0.88	1.05	0.96	1.08	1.44	1.20	1.47	1.11	1.27	1.00	0.91	1.28	0.89
F	0.61	1.09	1.66	0.80	0.70	0.96	1.22	0.50	0.91	0.80	0.76	0.97	1.45	1.37	0.75	0.89	0.90	0.92	0.82	0.91
G	1.15	0.80	1.05	1.32	1.11	1.62	1.25	1.00	0.94	1.10	1.06	1.17	1.53	1.03	1.21	1.14	1.15	0.99	0.56	1.21
H	1.24	0.62	1.23	1.10	0.89	0.72	0.95	0.85	0.90	1.02	0.87	0.94	1.88	0.94	1.09	1.31	1.09	0.85	1.01	0.88
I	1.12	0.59	0.96	0.90	0.83	1.02	0.71	0.63	0.67	0.82	0.72	0.94	1.12	0.96	0.70	0.87	0.88	0.62	0.57	0.91
K	0.82	0.89	0.89	1.15	1.00	1.09	1.12	0.74	0.77	0.78	0.93	1.18	1.05	1.07	0.80	1.13	0.99	0.95	0.87	0.90
L	0.74	0.74	1.31	0.75	0.84	1.13	1.09	0.77	1.03	0.81	0.87	0.98	1.40	1.00	0.81	1.06	0.91	0.89	0.65	0.81
M	0.74	0.42	0.93	0.85	0.70	0.98	1.15	1.12	0.71	0.82	0.79	0.98	1.10	1.04	0.92	0.81	0.77	0.68	0.81	0.76
N	1.13	0.65	1.42	1.05	1.18	1.25	0.99	1.06	0.91	0.96	1.01	1.20	1.51	1.05	1.06	1.15	1.00	0.94	0.55	1.08
P	1.45	0.70	1.45	1.66	1.57	1.36	1.67	1.25	1.43	1.39	0.98	1.17	2.08	1.59	1.32	1.50	1.55	1.45	1.06	1.17
Q	0.94	0.92	1.59	1.08	0.82	1.13	1.25	1.02	1.01	1.08	0.81	1.05	1.17	1.07	0.85	1.04	1.07	0.99	0.83	1.13
R	0.97	0.65	1.29	1.20	0.72	1.10	1.08	0.74	0.90	0.88	0.73	1.11	1.44	0.98	0.93	0.95	0.91	0.75	0.72	0.82
S	1.18	0.81	1.11	1.14	1.00	1.06	1.12	0.62	1.20	0.97	0.82	1.14	1.53	1.08	1.03	0.95	1.09	1.01	0.75	0.87
T	1.25	0.74	1.45	1.10	0.85	1.13	1.01	0.93	0.92	0.93	0.85	0.97	1.64	0.71	1.07	1.08	0.84	0.84	0.81	1.13
V	1.07	0.64	0.77	1.00	0.83	1.43	0.70	0.64	0.84	1.04	0.72	1.12	1.26	0.84	0.81	0.81	0.82	0.74	0.92	0.82
W	0.79	0.51	1.05	0.78	0.78	1.13	0.75	0.61	0.54	0.99	0.66	0.89	0.74	1.09	0.86	0.66	0.71	0.57	0.42	0.72
Y	0.86	0.88	1.05	0.98	0.90	1.13	0.81	0.60	0.78	0.69	0.91	1.03	1.61	0.85	0.84	0.96	0.87	0.87	0.86	0.87
Avg.	1.04	0.68	1.20	1.08	0.86	1.13	1.03	0.83	0.94	0.93	0.86	1.08	1.37	1.04	0.94	1.05	1.01	0.88	0.80	0.85

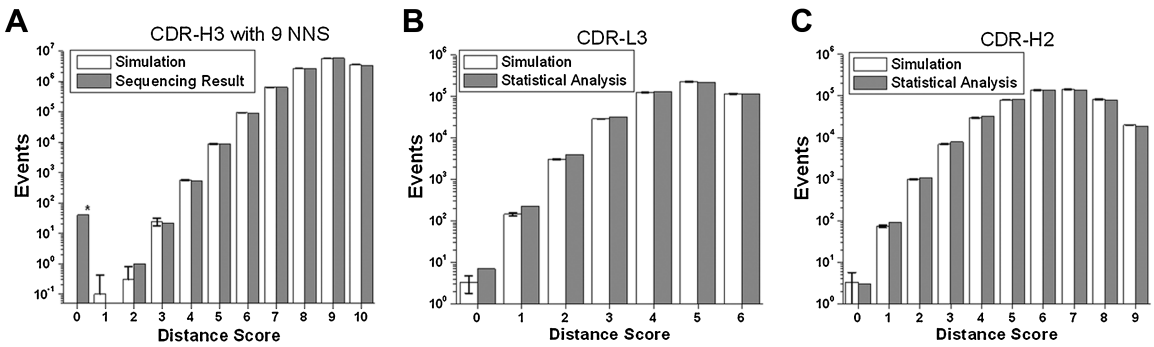

Figure 6. Comparisons of diversity distribution between simulation and statistical analysis of sequenced post-selection samples. The distance between two chosen CDR sequences was calculated by comparing position of the two CDR: if two sequences have the same amino acid at position i, then $d_i = 0$; otherwise $d_i = 1$. And the distance score is defined as, $D = \Sigma d_i$. The distance score profiles of all CDRs were generated by plotting the numbers of events having a particular distance score. The statistical analysis of sequenced samples was compared with simulation results of randomly generated CDR sequences using the designed codon randomization schemes. Ten simulation runs were performed for each CDR, and averages are shown with standard deviations as error bars. Sequence sample size was 5024 for CDR-H3 with nine NNS (**A**). Two thousand sequences were randomly chosen from the much larger sample pools of CDR-L3 (**B**) and CDR-H2 (**C**) to perform the analysis.

were identical; and (2) the duplicated sequences found in the pre-selection sample were different from those found in the post-selection sequence pool. These observations suggest that the duplicates were probably generated during the sequencing process. The phenomenon of repeated reads is well known for 454 sequencing (Gomez-Alvarez et al., 2009), and may arise during emulsion PCR when one water-in-oil droplet carries more than one bead, or during data collection when the optical signal from one well is recorded in an adjacent empty well (Briggs et al., 2007; Diehl et al., 2006). In our study duplicated sequences were not found in adjacent wells on the sequencing plate, and so we expect that the replicates were generated during emulsion PCR. In support of these conclusions, the $D = 1$ data match well with the simulation results (Fig. 6A).

Conclusions

Libraries encoding proteins with mutations at a variety of positions have so far been constructed using methods that require PCR amplification (Cobaugh et al., 2008; Feldhaus et al., 2003; Hoet et al., 2005; Knappik et al., 2000; Orlandi et al., 1989; Rajpal et al., 2005; Rothe et al., 2008; Silacci et al., 2005; Soderlind et al., 2000; Yin et al., 2008). However, PCR amplification can introduce a significant number of inadvertent mutations in constant regions since any errors in early amplification cycles are amplified during subsequent rounds (Cline et al., 1996; Pienaar et al., 2006). Also, gene assembly using partially degenerate primers can be difficult to optimize and may result in biased and/or reduced yields of full-length DNA (Kanagawa, 2003; Lueders and Friedrich, 2003; Polz and Cavanaugh, 1998; Sipos et al., 2007).

Advances in nucleotide chemistry have made possible the synthesis of ultra-long polynucleotides (up to ~220 bases) in large quantity and with very high purity. In this study, we show that large libraries can be generated easily by annealing relatively long oligonucleotides followed by gap filling by T4 DNA polymerase. This methodology results in a high yield of full-length, purified gene product within hours. Primarily due to its very active 3′–5′ exonuclease activity (Rehakrantz et al., 1991), T4 DNA polymerase displays excellent fidelity (estimated at less than one mis-incorporation per 10^7 residues (Kunkel et al., 1984)), thus allowing the generation of libraries having a very low rate of unwanted mutations in regions not subjected to randomization. In contrast, even the highest fidelity thermophilic DNA polymerase used for PCR, such as *Pfu* or *Vent*, display errors rates (1.3 and 2.8×10^{-6}) more than an order of magnitude greater (Cline et al., 1996). In addition, the present method does not require a template, with both strands of the annealed pairs elongating at the same time. Overall, the nature of elongation by high fidelity polymerase without amplification give this method unique advantages comparing to PCR-based manners, in terms of reserving designed diversity and minimizing bias among variable regions, and high accuracy in the constant regions.

We used this methodology to generate novel synthetic libraries of V_H and V_κ genes for the isolation of antibodies by combinatorial screening methods. For therapeutic applications it is generally desirable to employ libraries that do not contain unintended mutations in the FRs which could result in immunogenicity or alternatively affect the stability or folding of the variable domains. While mutations in the frameworks regions also occur in high affinity antibodies produced by the mammalian immune system, following somatic hypermutation of B cells in the germinal centers, those mutations are subject to negative selection during subsequent B cell development. Since this is not the case for antibodies isolated from synthetic libraries, it is preferable to employ libraries that comprise of frameworks regions identical to the germline sequences. The higher fidelity inherent to library construction by gap filling using T4 polymerase compared to PCR-based methods may

significantly reduce the frequency of clones containing such unintended mutations. While for this work we have employed four oligos to synthesize variable domain genes between 340 and 390 bp, we have also constructed V genes using only two oligonucleotides albeit in that case library quality was lower due to the higher number of deletions introduced by chemical synthesis (data not shown). Here we chose to introduce the assembled genes into expression vectors by ligation, but this step can be eliminated using recombineering strategies (Hartley et al., 2000). Successful evolution of high affinity antibodies often requires further diversification after selection to improve affinity. While randomization of the entire V gene and library screening has been used successfully in numerous academic affinity maturation studies, for therapeutic antibodies the introduction of mutations in the FRs is a concern because of the potential that such amino acid substitutions can be immunogenic. Therefore, for therapeutic antibody development affinity and stability engineering of the antibody is typically achieved by mutagenizing only the CDR regions. The library construction method described here could also be exploited to rapidly re-randomize desired CDRs of selected clones.

High-throughput DNA sequencing has been widely used to study genomic DNA, and recently the naïve antibody repertoires of zebrafish and human were analyzed by a HTS approach (Glanville et al., 2009; Weinstein et al., 2009). In our study, HTS was employed to characterize the synthetic human antibody libraries generated by annealing/gap filling, before and after in-frame selection using β-lactamase fusions. Statistical results of CDR sequences (~48,000 on average) in the V_H and V_κ libraries revealed that: (i) The expected amino acid diversity is attained prior to selection for expression; (ii) the in-frame selection depleted stop codons as expected but did not alter the distribution of other amino acids except from Cys whose abundance was significantly reduced, especially in CDR-L3; (iii) excellent diversity was obtained in randomized regions and was essentially indistinguishable from the theoretical diversity determined The post-selected V_H and V_κ genes have been incorporated in both scFv and full-length IgG libraries for screening by phage or bacterial display, respectively. Furthermore, the methodology described herein could be readily employed of other protein engineering applications requiring the diversification of multiple non-consecutive regions in a protein.

We acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for postdoctoral fellowship to X.G. We are grateful to M. Jack Borrok and Sai Reddy for comments on the MS. This work was supported by a grant from the Clayton Foundation.

References

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M, Paabo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104(37):14616–14621.

Clark M. 2000. Antibody humanization: A case of the 'Emperor's new clothes'? Immunol Today 21(8):397–402.

Cline J, Braman JC, Hogrefe HH. 1996. PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 24(18): 3546–3551.

Cobaugh CW, Almagro JC, Pogson M, Iverson B, Georgiou G. 2008. Synthetic antibody libraries focused towards peptide ligands. J Mol Biol 378(3):622–633.

Diehl F, Li M, He YP, Kinzler KW, Vogelstein B, Dressman D. 2006. BEAMing: Single-molecule PCR on microparticles in water-in-oil emulsions. Nat Methods 3(7):551–559.

Ewert S, Huber T, Honegger A, Pluckthun A. 2003. Biophysical properties of human antibody variable domains. J Mol Biol 325(3):531–553.

Feldhaus MJ, Siegel RW, Opresko LK, Coleman JR, Feldhaus JMW, Yeung YA, Cochran JR, Heinzelman P, Colby D, Swers J, et al. 2003. Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. Nat Biotechnol 21(2):163–170.

Fellouse FA, Esaki K, Birtalan S, Raptis D, Cancasci VJ, Koide A, Jhurani P, Vasser M, Wiesmann C, Kossiakoff AA, Koide S, Sidhu SS. 2007. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. J Mol Biol 373(4):924–940.

Glanville J, Zhai WW, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GMR, Cox D, Rajap A, Pons J. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA 106(48): 20216–20221.

Hartley JL, Temple GF, Brasch MA. 2000. DNA cloning using in vitro site-specific recombination. Genome Res 10(11):1788–1795.

Hayashi N, Welschof M, Zewe M, Braunagel M, Dubel S, Breitling F, Little M. 1994. Simultaneous mutagenesis of antibody CDR regions by overlap extension and PCR. Biotechniques 17(2):310.

Hayhurst A, Happe S, Mabry R, Koch Z, Iverson BL, Georgiou G. 2003. Isolation and expression of recombinant antibody fragments to the biological warfare pathogen Brucella melitensis. J Immunol Methods 276(1–2):185–196.

Hecker KH, Rill RL. 1998. Error analysis of chemically synthesized polynucleotides. Biotechniques 24(2):256–260.

Hoet RM, Cohen EH, Kent RB, Rookey K, Schoonbroodt S, Hogan S, Rem L, Frans N, Daukandt M, Pieters H, van Hegelsom R, Neer NC, Nastri HG, Rondon IJ, Leeds JA, Hufton SE, Huang L, Kashin I, Devlin M, Kuang G, Steukers M, Viswanathan M, Nixon AE, Sexton DJ, Hoogenboom HR, Ladner RC. 2005. Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity. Nat Biotechnol 23(3):344–348.

Hoogenboom HR. 2005. Selecting and screening recombinant antibody libraries. Nat Biotechnol 23(9):1105–1116.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8(7):9.

Kanagawa T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96(4):317–323.

Knappik A, Ge LM, Honegger A, Pack P, Fischer M, Wellnhofer G, Hoess A, Wolle J, Pluckthun A, Virnekas B. 2000. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. J Mol Biol 296(1):57–86.

Kunkel TA, Loeb LA, Goodman MF. 1984. On the fidelity of DNA replication. The accuracy of T4 DNA polymerases in copying phi X174 DNA *in vitro*. J Biol Chem 259(3):1539–1545.

Lee CV, Liang WC, Dennis MS, Eigenbrot C, Sidhu SS, Fuh G. 2004. High-affinity human antibodies from phage-displayed synthetic fab libraries with a single framework scaffold. J Mol Biol 340(5):1073– 1093.

Liang WC, Dennis MS, Stawicki S, Chanthery Y, Pan Q, Chen YM, Eigenbrot C, Yin JP, Koch AW, Wu XM, Ferrara N, Bagri A, Tessier-Lavigne M, Watts RJ, Wu Y. 2007. Function blocking antibodies to neuropilin-1 generated from a designed human synthetic antibody phage library. J Mol Biol 366(3):815–829.

Lueders T, Friedrich MW. 2003. Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. Appl Environ Microbiol 69(1):320–326.

Lutz S, Fast W, Benkovic SJ. 2002. A universal, vector-based system for nucleic acid reading-frame selection. Protein Eng 15(12):1025–1030.

Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. 1991. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. J Mol Biol 222(3):581–597.

Martin ACR. 1996. Accessing the Kabat antibody sequence database by computer. Proteins Struct Funct Genet 25(1):130–133.

Mazor Y, Van Blarcom T, Iverson BL, Georgiou G. 2008. E-clonal antibodies: Selection of full-length IgG antibodies using bacterial periplasmic display. Nat Protoc 3(11):1766–1777.

Mazor Y, Van Blarcom T, Mabry R, Iverson BL, Georgiou G. 2007. Isolation of engineered, full-length antibodies from libraries expressed in Escherichia coli. Nat Biotechnol 25(5):563–565.

Orlandi R, Gussow DH, Jones PT, Winter G. 1989. Cloning immunoglobulin variable domains for expression by the polymerase chain reaction. Proc Natl Acad Sci USA 86(10):3833–3837.

Pienaar E, Theron M, Nelson M, Viljoen HJ. 2006. A quantitative model of error accumulation during PCR amplification. Computat Biol Chem 30(2):102–111.

Polz MF, Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64(10):3724–3730.

Puigbo P, Guzman E, Romeu A, Garcia-Vallve S. 2007. OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res 35:W126–W131.

Rajpal A, Beyaz N, Haber L, Cappuccilli G, Yee H, Bhatt RR, Takeuchi T, Lerner RA, Crea R. 2005. A general method for greatly improving the affinity of antibodies by using combinatorial libraries. Proc Natl Acad Sci USA 102(24):8466–8471.

Rehakrantz LJ, Stocki S, Nonay RL, Dimayuga E, Goodrich LD, Konigsberg WH, Spicer EK. 1991. DNA polymerization in the absence of exonucleolytic proofreading: In vivo and in vitro studies. Proc Natl Acad Sci USA 88(6):2417–2421.

Rothe C, Urlinger S, Lohning C, Prassler J, Stark Y, Jager U, Hubner B, Bardroff M, Pradel I, Boss M, Bittlingmaier R, Bataa T, Frisch C, Brocks B, Honegger A, Urban M. 2008. The human combinatorial antibody library HuCAL GOLD combines diversification of all six CDRs according to the natural immune system with a novel display method for efficient selection of high-affinity antibodies. J Mol Biol 376(4):1182–1200.

Seehaus T, Breitling F, Dubel S, Klewinghaus I, Little M. 1992. A vector for the removal of deletion mutants from antibody libraries. Gene 114(2):235–237.

Sidhu SS, Fellouse FA. 2006. Synthetic therapeutic antibodies. Nat Chem Biol 2(12):682–688.

Sidhu SS, Li B, Chen Y, Fellouse FA, Eigenbrot C, Fuh G. 2004. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. J Mol Biol 338(2):299–310.

Silacci M, Brack S, Schirru G, Marlind J, Ettorre A, Merlo A, Viti F, Neri D. 2005. Design, construction, and characterization of a large synthetic human antibody phage display library. Proteomics 5(9):2340–2350.

Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol 60(2):341–350.

Soderlind E, Strandberg L, Jirholt P, Kobayashi N, Alexeiva V, Aberg AM, Nilsson A, Jansson B, Ohlin M, Wingren C, Danielsson L, Carlsson R, Borrebaeck CA. 2000. Recombining germline-derived CDR sequences for creating diverse single-framework antibody libraries. Nat Biotechnol 18(8):852–856.

Gomez-Alvarez V, Teal TK, Schmidt TM. 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME J 3(11):1314–1317.

Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. Science 324(5928):807–810.

Yin CC, Ren LL, Zhu LL, Wang XB, Zhang Z, Huang HL, Yan XY. 2008. Construction of a fully synthetic human scFv antibody library with CDR3 regions randomized by a split-mix-split method and its application. J Biochem 144(5):591–598.

Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, Schroeder Jr HW, Kirkham PM. 2003. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. J Mol Biol 334(4):773–749.