

*Antibodies* **2015**, *4*, 88-102; doi:10.3390/antib4020088

OPEN ACCESS

*antibodies*

ISSN 2073-4468

[www.mdpi.com/journal/antibodies](http://www.mdpi.com/journal/antibodies)

Article

## Codon-Precise, Synthetic, Antibody Fragment Libraries Built Using Automated Hexamer Codon Additions and Validated through Next Generation Sequencing

Laura Frigotto <sup>1</sup>, Matthew E. Smith <sup>1</sup>, Christopher Brankin <sup>1,†</sup>, Ashni Sedani <sup>1</sup>, Simon E. Cooper <sup>1</sup>, Nisha Kanwar <sup>1</sup>, Daniel Evans <sup>1</sup>, Stanislava Svobodova <sup>1,‡</sup>, Claudia Baar <sup>1</sup>, Jacob Glanville <sup>2</sup>, Christopher G. Ullman <sup>1,\*</sup> and Anna V. Hine <sup>3,\*</sup>

<sup>1</sup> Isogenica Ltd., The Mansion, Chesterford Research Park, Little Chesterford, Essex, CB10 1XL, UK; E-Mails: [Laura.Frigotto@isogenica.com](mailto:Laura.Frigotto@isogenica.com) (L.F.); [Matthew.Smith@isogenica.com](mailto:Matthew.Smith@isogenica.com) (M.E.S.); [brankinc@MedImmune.com](mailto:brankinc@MedImmune.com) (C.B.); [Ashni.Sedani@isogenica.com](mailto:Ashni.Sedani@isogenica.com) (A.S.); [Simon.Cooper@isogenica.com](mailto:Simon.Cooper@isogenica.com) (S.E.C.); [Nisha.Kanwar@isogenica.com](mailto:Nisha.Kanwar@isogenica.com) (N.K.); [Daniel.Evans@isogenica.com](mailto:Daniel.Evans@isogenica.com) (D.E.); [ss@hc-d.co.uk](mailto:ss@hc-d.co.uk) (S.V.); [Claudia.Baar@isogenica.com](mailto:Claudia.Baar@isogenica.com) (C.B.)

<sup>2</sup> Distributed Bio Inc, 660 4th St, Suite 491, San Francisco, CA 94107, USA; E-Mail: [jake@distributedbio.com](mailto:jake@distributedbio.com)

<sup>3</sup> School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK

<sup>†</sup> Current address: MedImmune, Milstein Building, Granta Park, Cambridge, CB21 6GH, UK.

<sup>‡</sup> Current address: Healthcare Diagnostics Ltd., Rossington Place, Graphite Way, Hadfield, Glossop, Derbyshire SK13 1QG, UK.

\* Authors to whom correspondence should be addressed;

E-Mails: [Chris.Ullman@paratopix.com](mailto:Chris.Ullman@paratopix.com) (C.G.U.); [a.v.hine@aston.ac.uk](mailto:a.v.hine@aston.ac.uk) (A.V.H.);  
Tel.: +44-121-204-3961 (A.V.H.); Fax: +44-121-204-5060 (A.V.H.).

Academic Editor: Dimiter S. Dimitrov

Received: 5 March 2015 / Accepted: 11 May 2015 / Published: 15 May 2015

---

**Abstract:** We have previously described ProxiMAX, a technology that enables the fabrication of precise, combinatorial gene libraries via codon-by-codon saturation mutagenesis. ProxiMAX was originally performed using manual, enzymatic transfer of codons via blunt-end ligation. Here we present Colibra™: an automated, proprietary version of ProxiMAX used specifically for antibody library generation, in which double-codon hexamers are transferred during the saturation cycling process. The reduction in process complexity, resulting library quality and an unprecedented saturation of up to 24 contiguous

codons are described. Utility of the method is demonstrated via fabrication of complementarity determining regions (CDR) in antibody fragment libraries and next generation sequencing (NGS) analysis of their quality and diversity.

**Keywords:** Colibra; ProxiMAX randomization; saturation mutagenesis; CDR; single domain antibody; camelid antibody; antibody engineering

---

## 1. Introduction

The ability to construct exquisitely-designed, synthetic protein libraries is a powerful tool in protein engineering, especially for randomizing complementarity determining regions (CDRs) in antibody binding sites. By mutagenizing the CDRs, novel high-affinity antibodies can be isolated from man-made systems using molecular display and evolution methods [1–4]. However, the performance of synthetic libraries has not always been optimal since early methods of mutagenesis involved saturation of the encoding gene with degenerate codons such as NNK (where N = A/C/G/T & K = G/T) [5]. Such saturation necessarily encodes all 20 amino acids in ratios dictated by the genetic code. Moreover, this saturation severely limits the functional diversity of such libraries by encoding amino acid composition that is unable to fold natively, and leads to premature termination through the encoding of a stop codon, which in turn can lead to non-functional proteins [5–7]. Several solutions to reduce redundancy in the code and exclusion of termination signals has led to inventions such as the 22c trick (all 20 amino acids are encoded by 22 codons through degenerate oligonucleotides), offering near non-degenerate randomization [8]. These degenerate libraries still have the potential limitations of encoding problematic residues such as cysteine and methionine; an issue addressed by MDC-Analyzer [9], which controls encoded amino acids, but not their relative proportions and thus cannot be adapted to finely tune the amino acid composition so that only desired ratios of functional residues are included.

Therefore, to achieve optimal diversity of functional sequence space within a library, a capacity to define precise amino acid composition, on a position by position and a loop by loop basis is desired. Analysis of *in-vivo* antibody repertoires have identified strong positional bias of amino acids tolerated by antibody folding [3,5,6,10]. Additionally, the number, identity and position of residues favored to make antigen contacts in both V<sub>H</sub> and V<sub>L</sub> domains varies both according to the type of antigen (protein, peptide or hapten), thus providing information that can be used to design libraries that have a greater proportion of functional members or even to design target-specific antibody libraries [3]. Previous synthetic methods (TRIM, SlonoMAX [11] and ProxiMAX [7]) have developed approaches to control amino acid identities and frequencies on a per-position basis. These all create DNA fragments in a codon-by-codon manner, though via different means. TRIM functions by adding three bases at a time, using trinucleotide phosphoramidites, to a growing single strand of synthetic DNA. In contrast, SlonoMAX [11] ligates up to 4096 sticky-ended “anchors” to up to 64 “splinkers” (both are hairpin oligonucleotides with 3-base, single-stranded overhangs) together. The ligated product is then cleaved with a TypeII S restriction enzyme that leaves a new 3-base overhang on the splinker [11]. In this manner, a single-stranded codon is effectively added to the splinker in each cycle. Finally, ProxiMAX uses a blunt-ended, double-stranded approach. Up to 20 blunt-ended, double stranded oligonucleotide “donor” sequences

are ligated to the blunt-ended, double-stranded “acceptor”. The ligated product is then cleaved with a Type IIS restriction enzyme that leaves a blunt end, so that the double-stranded codon is effectively transferred from the donor to the acceptor and the cycle is repeated (Figure 1) [12]. By using three rotating sets of 20 donors, exquisite control of codon identity and frequency can be obtained [7].

However, such methods had not traditionally attempted to eliminate higher-order amino acid motifs that can form major liabilities for protein development, such as deamidation motifs NG, NS and NA, isomerization motif DG, proteolytic cleavage motifs KK, RK and KR, N-linked glycosylation sites N<sup>[^P]</sup>[ST] and plastic binding motif WxxW. Although Tiller *et al.* [13] have recently alluded to the removal of potential post-translational modification sites (including deamidation, isomerization, protease cleavage and oxidation sites plus methionine) via “diversity” within the CDR-H1 and CDR-H2 regions of their synthetic human Fab library, they do not describe how this removal was achieved and the relevant gene fragments do not appear to have been made exclusively by SlonoMAX technology. Ideally, whatever would be the ultimate use of the protein library, the protein engineer should be able saturate codons to encode a *selected* choice of amino acids, in *selected* proportions at *selected* positions, while selectively excluding the appearance of higher-order motifs as part of the saturation process. This would allow the desirable elimination of amino acids such as cysteine and methionine (prone to oxidation [14]), as well as complex liabilities such as acid hydrolysis, deamidation, proteolytic cleavage and glycosylation that impede downstream manufacture of the encoded antibodies [4], in addition to minimizing the designed library size *and* empowering the scientist to create new repertoires or mimic the sequence fitness landscapes elected by evolutionary forces [6].

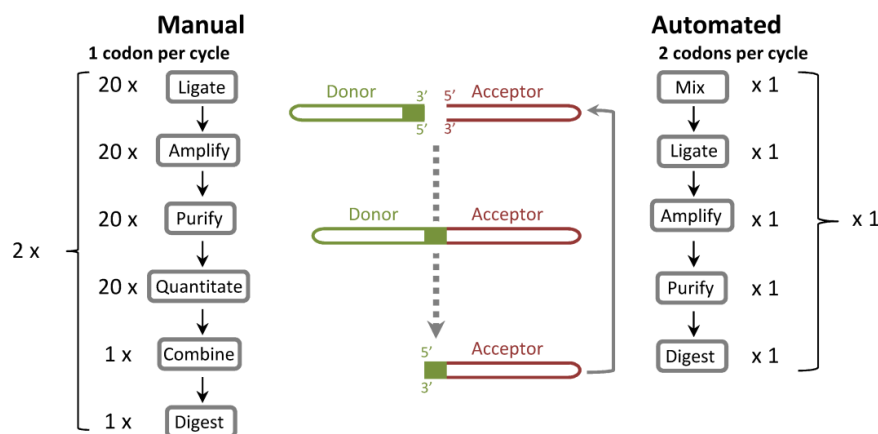
Previously we have described the ProxiMAX technique which can dictate precise additions of single codons (trimer nucleotides, “trimers”) to form a randomized CDR3 region of a V<sub>H</sub> domain. Whilst this technology can be operated at the bench with little investment in equipment, it can be cumbersome and labor intensive as a procedure at scale, hence, simplification and automation of the technique have been pursued to improve the high throughput manufacture of libraries. Additionally, efficiency and performance enhancements could be achieved by using hexamer nucleotides (double codons, “hexamers”) rather than single trimer sets. In so doing, this would introduce two sequential randomized amino acids simultaneously into the protein chain with the precise control of the amino acid ratios offered by the ProxiMAX technique. Van den Brulle *et al.* described a ligation technique, SlonoMAX, for library synthesis in which hexamer additions were used, however this was solely as a requirement to anticipate ligation of the next codon addition in the sticky ended ligations, which used 3-base pair overhangs. Therefore, only single codon randomizations were achieved through each addition in the SlonoMAX method [11]. Herein, we describe the incorporation of complete hexamer cocktails, deliberately eliminating liabilities within hexamer cocktails by simple exclusion, and possibly from across junction sites by design.

In this study, we use ProxiMAX hexamer randomization to deliver exquisite control of library fabrication. Due to the complex nature of these mixtures and additions, we proceed from ligations and hand-mixing of discrete ligated hexamer products through to automation of mixed pools of hexamer components followed by pool-ligation. We exemplify the utility of ProxiMAX’s hexamer additions for fabrication in multiple settings, including scFv CDR-L3 cassettes and camelid V<sub>HH</sub> CDR3-H3 cassettes. In general the CDR3 loops from V<sub>HH</sub> domains are longer than those of CDR-H3s from human and mouse repertoires [3,4,15]. Fabrication of such extreme loop lengths present further technical challenges which

would have been difficult to address using the single trimer additions of ProxiMAX. We have also integrated next generation sequencing (NGS) techniques in order to monitor the library quality at critical steps of fabrication [10].

## 2. Results and Discussion

ProxiMAX randomization is a mutagenesis technique that can saturate at defined positions within a protein with any specified mixture of codons in any specified ratio. Gene fragments are built one codon at a time, via a process of enzymatic saturation cycling [7] (Figure 1). In our original publication of the methodology, we described the manual construction of a model CDR-H3 library, in which we were able to exclusively encode mixtures of selected amino acids, in user-defined, 5% incremental ratios (for example, CDR-H3 loop Kabat position 100B: 20% Asp; 20% Ser; 20% Tyr; 10% Gly; 10% Thr; 10% Val; 5% Ala; 5% Arg). In every saturation cycle of this manual process, up to 20 donor sequences were individually ligated, amplified, purified and quantified, then combined and digested. The methodology was ideal for constructing the highest quality libraries as exceptional agreement between design specifications and observed library composition was previously demonstrated, but it was relatively expensive and labor intensive as a commercial process.



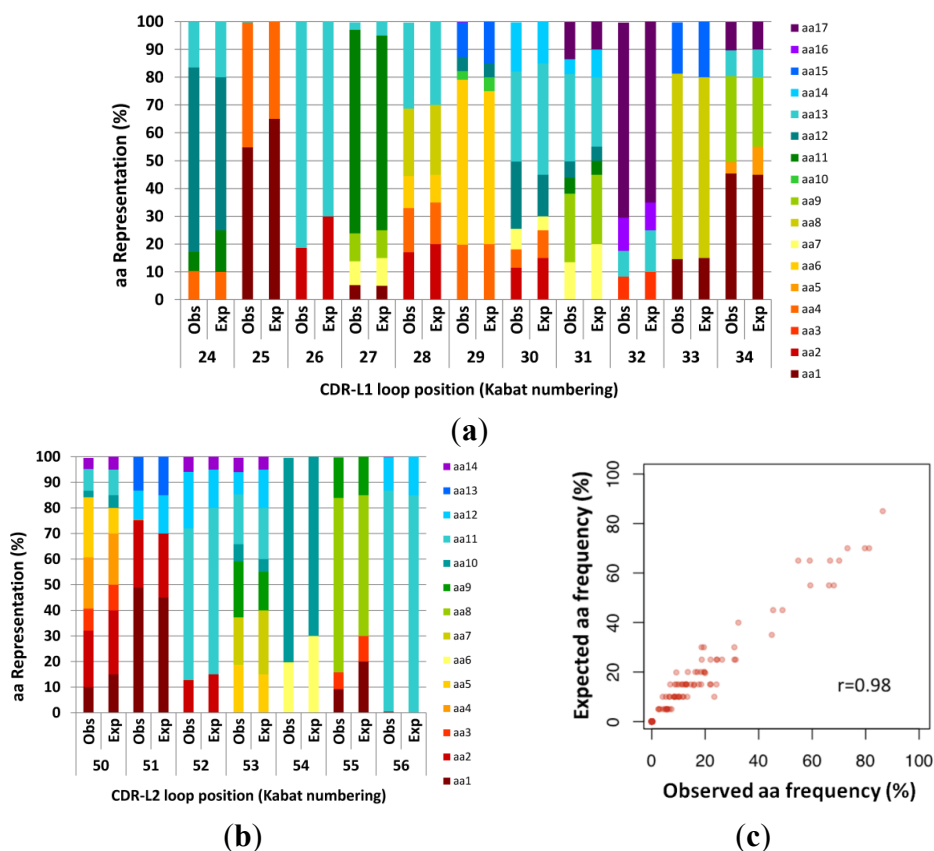
**Figure 1.** Comparison of ProxiMAX and automated, hexamer-based ProxiMAX (Colibra™) processes. In the manual process, codons are ligated separately, amplified, purified, quantified, mixed into a pool of ligated products and then digested with *MlyI*. This procedure transfers the codon from a donor molecule to an acceptor library element. In an automated hexamer assembly, all hexamer components are premixed, ligated as a pool, PCR amplified, purified and then cleaved with *MlyI*. One hexamer cycle is equivalent to two manual trimer cycles.

### 2.1. Colibra Development: Manual Hexamer Method Validation

To determine process feasibility, scFv light chain CDR regions were constructed using manual hexamer addition. Initially, as in the published ProxiMAX protocol [7], the hexamers were ligated separately and individual components were purified and then mixed at the intended ratios to generate a library that reflected the design of the CDRs. Individual ligations were undertaken owing to previous observed biases caused by sequence preferences of T4 ligase during trimer addition, such as for the His codon, CAT [7]. We assumed that similar biases might be present at the ligation junctions of the

hexamers which could skew the library synthesis towards sequences for which T4 ligase had a greater preference. A semi-minimalist design was chosen that would be amenable to manual pipetting, with codon frequencies rounded to the nearest 5%.

The longer light chain CDRs were fabricated using the bi-directional multi-cassette approach previously described [7], whereby a randomized region can be constructed as multiple cassettes that are then joined together to obtain the desired length. A major advantage of the hexameric approach, when compared to the single codon/trimeric methodology, is the capacity to achieve greater lengths whilst decreasing the number of saturation cycling steps required. Moreover, since the quality of the final product is related to the repeating cycles of additions, decreasing the number of steps by the use of hexamers has the potential of improving the quality of the library by reducing the accumulation of spurious ligation products and codon deletions or base deletions resulting from oligonucleotide impurities. Base deletions in the trimer stocks were measured to be  $1.30\% \pm 0.42\%$  whereas those for hexamers were  $1.16\% \pm 0.13\%$ . For CDRs of shorter length, there is an advantage in using hexamers since the number of cassettes can be minimized, reducing synthesis time. Furthermore, when hexamer cassettes are used, much longer CDR loops can be attained.



**Figure 2.** Expected (Exp) *versus* observed (Obs) amino acid frequencies in complementarity determining regions (CDR) fragment libraries created using manual hexamer additions. Codon frequencies were determined by Illumina Miseq analysis of amplicons of expected length. Designed (expected) codons ranged from 5% to 85% per position. (a) CDR-L1. (b) CDR-L2. (c) Expected design frequencies plotted *versus* Observed frequencies calculated from next generation sequencing (NGS) ( $n = 130399$ ). The resulting library was highly correlated to the design (Pearson's  $r = 0.984$ ).

Whilst the synthesis of our designed CDR-L2 did not require multiple cassettes, we fabricated CDR-L1 by adding 3 hexamers on one cassette (6 codons), and 2 hexamers and a trimer (5 codons) on a second cassette, in order to obtain the desired 11 codon loop. The two cassettes were then ligated together using the same approach as during synthesis, as described in the Experimental Section. The fabrication of scFv light chain CDR-L1 and CDR-L2 demonstrate the fine control of hexamer codon additions across the CDR length with all additions being maintained close to the ratios that were designed (Figure 2).

In theory, hexamers should result in greater library purity when compared with multiple trimer additions, since impurities accumulate per addition. Comparison with a scFv CDR-H2 loop built by trimer addition (unpublished data) indeed suggests some improved product purity (cf. 88.3% correct length from trimer addition with 92.1% by using hexamers). To generate a scFv library, CDR-L1 and CDR-L2 fragments described in Figure 2 were combined with exemplar CDR-L3 loops that had been previously fabricated. The resulting light chain CDR segments were then ligated to framework regions to generate V<sub>L</sub> domains and the percentage of products with the expected length was analyzed (Table 1). Finally, the V<sub>L</sub> domains were ligated to a V<sub>H</sub> library previously constructed using ProxiMAX trimer methodology, to generate complete scFv libraries.

**Table 1.** Length accuracy of synthetic scFv loops and domains. Light chain CDR loops were analysed by Illumina Miseq sequencing (n = 351021 for CDR-L1 and 345614 for CDR-L2) and the results verified by Sanger sequencing (n = 102). CDR-L1 and CDR-L2 were then combined with exemplar CDR-L3 domains encoding loop lengths of both 9 and 10 aa's, to generate light chain libraries V<sub>L</sub> 3-9 and V<sub>L</sub> 3-10 respectively. The percentage of functional (in frame) fully-assembled light-chain V<sub>L</sub> sequences was examined by Sanger sequencing.

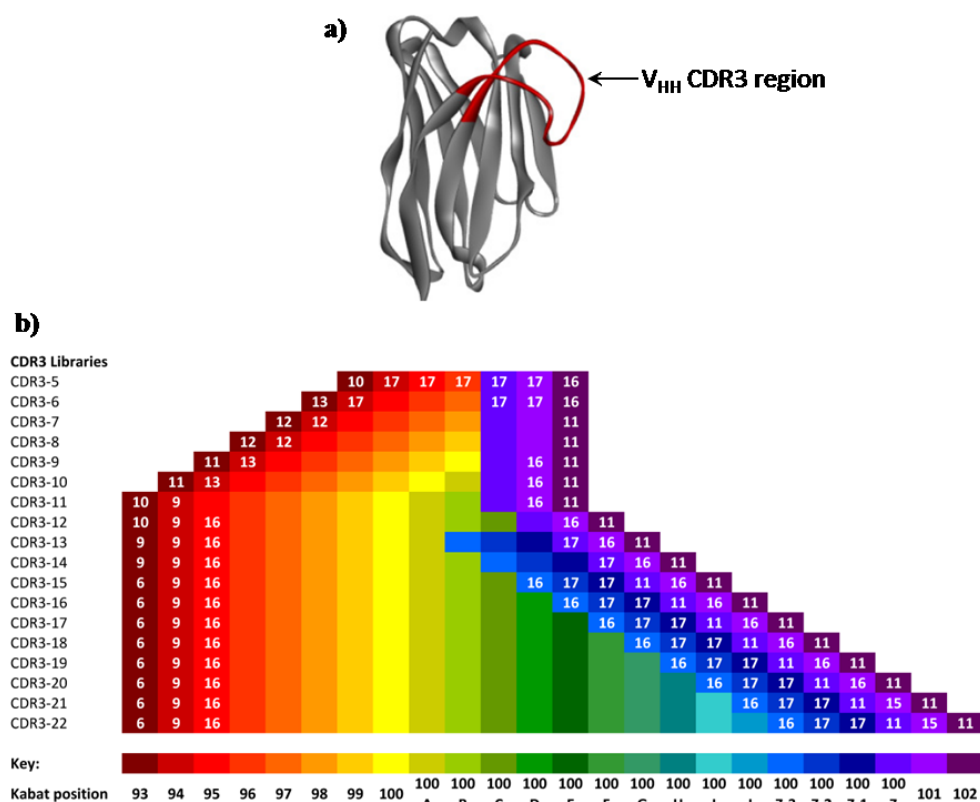
Domain	Correct length	n-1	n-3	Sequencing methodology
CDR-L1	92%	7%	0.5%	Illumina NGS
	92%	8%	0.0%	Sanger
CDR-L2	96%	3%	0.5%	Illumina NGS
	98%	2%	0.0%	Sanger
V <sub>L</sub> 3-9	77%	n/a	n/a	Sanger
V <sub>L</sub> 3-10	77%	n/a	n/a	Sanger

## 2.2. Automated Hexamer Method Validation

Having determined that hexamer addition was a viable process, we next turned attention to automation. In order to make the process more economical, manageable and more flexible in terms of codon use and frequency, we developed a mixed pot procedure to reduce both the number of process steps and the requirement for manual manipulation of the products within each cycle. Specifically, liquid-handling robots were utilized to enable a pool of 400 hexameric donor sequences to be employed (as compared with 20 trimeric donors) at each cycle and 3 sets of such oligonucleotides (total 1200 oligonucleotides) were used over three sequential ligations, which also required three sets of post-ligation PCR recovery oligonucleotides. Whilst we used 20 standard codons, one coding for each amino acid, others have been tested and this provides an additional advantage over chemical synthesis methods where the complete codon set has not been manufactured. The expense in reagents was compensated by improved throughput in the process: for the randomization of two consecutive codons, this constituted a

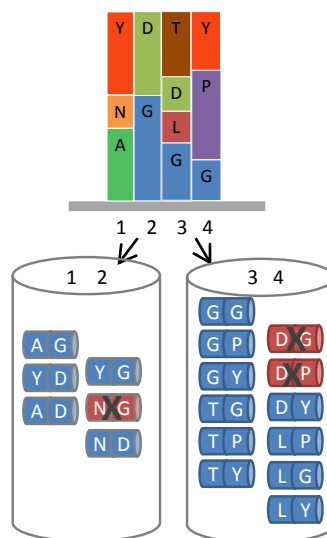
reduction from 164 to just 5 steps (albeit with up to 400 automated pipetting actions for each randomized hexamer position, Figure 1).

To ‘pressure test’ the relative quality of this automated hexamer process, we sought to challenge the technology by synthesising multiple, long V<sub>HH</sub> CDR loops at user-defined mixtures of codons in sometimes less than 1% ratios. The design was based on camel, llama and alpaca repertoires and from bioinformatical analyses of next generation sequencing data sets. Camelid CDR3s are generally longer than those of other characterized vertebrate repertoires (with the exception of bovine CDR3 domains derived from a specialized DH2 region) and potentially have different target preferences from shorter CDR3s [16]. Therefore a range of V<sub>HH</sub> CDR3 domains, with loop lengths of between 7 and 24 saturated residues in our design (5–22 amino acids within CDR3 region 95–102 as defined by Kabat [17]) were created (Figure 3).



**Figure 3.** (a) Ribbon representation of an exemplar llama V<sub>HH</sub> domain (pdb1I3V<sup>16</sup>), showing the extended CDR3 domain targeted for saturation mutagenesis. (b) Illustration of the Kabat positions (black font) and numbers of encoded amino acids at each Kabat position (white font, <18 encoded amino acids) within each library segment. Absence of a white number indicates that all 18 amino acids (no Cys or Met) were encoded.

In the synthesis of these loops, it was possible to remove codons for methionine and cysteine from all saturated positions. It was also possible to eliminate the encoding of the undesirable amino acid combinations NG, NS, NA, DG, DP, DS (which render the protein product susceptible to hydrolysis, deamidation and isomerisation [4]) within hexanucleotide additions (Figure 4), by simple omission of relevant hexanucleotide donors (Note that such elimination *between* hexanucleotide additions was not attempted in the present study but is achievable, if required, via a split-pot synthetic approach).



**Figure 4.** Removal of liabilities from hexameric ProxiMAX. A hypothetical gene region requiring 4 positions of saturation mutagenesis (1 to 4, top diagram), each position having varying codon number, identity and frequency, can be fabricated by automated hexameric ProxiMAX by sequential addition of two pools of MAXMAX codons (1–2 and 3–4, bottom diagram) containing the required mixture of double-codons at a defined percentage, dictated by the specific library design. Any unfavorable codon pair can be selectively removed from the mixes, increasing the functionality of the library.

In order to achieve precision in the mixing and addition of the hexamers, liquid handling robots were employed to mix dilutions of annealed oligonucleotide stocks with individual volumes ranging from 1 to 96  $\mu$ L. Automation equipment was calibrated for each liquid class and where a concentrated oligonucleotide stock was required, the mix was performed in duplicate and combined into a stock solution of hexamers to reduce pipetting inaccuracies and mistakes in the aspiration or dispense steps.

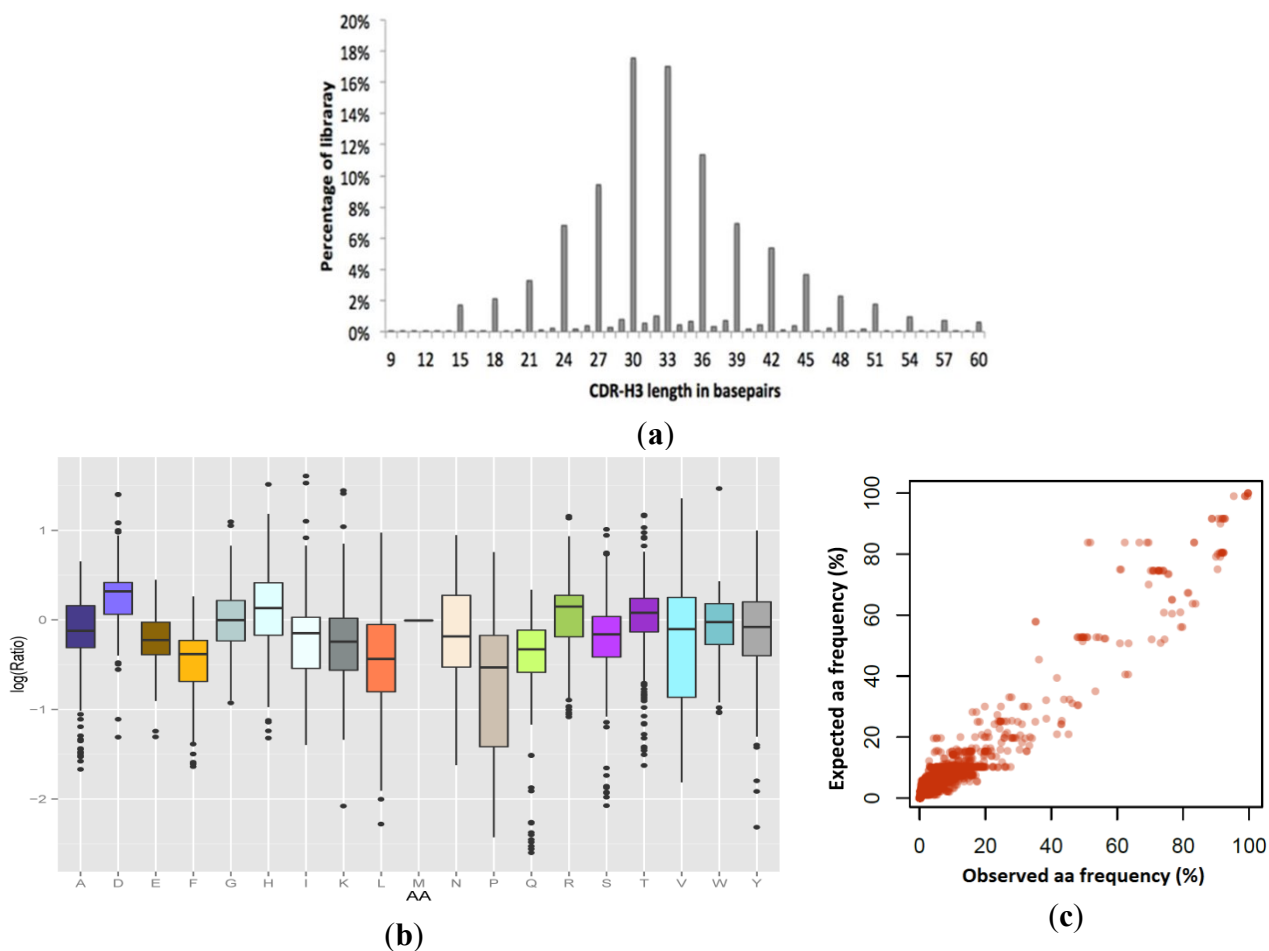
Prior to their use in the automated ProxiMAX, the hexameric mixes were analysed by NGS to verify both the presence and the proportion of all intended donors (up to 400 per mix). Acceptance criteria dictated that all desired hexameric oligos were present, but excluding those that were not designed. However, adjustments for instrumentation error rates, batch-to-batch and sampling variations, as well as allowances for some deviation of the observed frequencies from the expected levels, similar to those described in Table 1 determined a ‘QC pass’. On average, greater than 94% of hexameric components in each mix complied with such criteria, and those mixes that fell below this level were repeated.

Variable regions CDR3-5 to CDR3-22 (Figure 3b) were synthesized individually. Where longer CDR3 regions were required, these were generated in two cassettes and subsequently ligated together (CDR3-13 to CDR3-22). The 18 individual libraries were then combined in a Gaussian distribution of length and ligated to CDR1 and CDR2 regions built similarly, to generate a library of  $V_{HH}$  domains. The distribution of the combined pool, the encoded amino acid identity and the fidelity of designed ratios in the saturated positions were examined by Illumina NGS and subsequent analyses (Figure 5). The CDR3 loop length in this library was skewed towards the distribution of mouse and human CDR3s for a comparative study of this library against equivalent targets from other synthetic libraries, (although alternative distributions could be prepared). Furthermore, a combination analysis of the amino acid



incorporation efficiencies at 255 independent positions in the V<sub>HH</sub> library was performed. The expected incorporation of each amino acid in the resulting library design was compared to the observed incorporation and the fidelity per each amino acid found to be high (Pearson's  $r = 0.963$ ).

Overall, the synthesis yielded CDR3 regions that were 92% pure (by size) and 99.7% of all amino acid additions passed the QC criteria (Table 2). In virtually all positions for all amino acids in the design, the designed frequency was achieved within our QC boundaries and each loop was synthesized to greater than 85% purity.



**Figure 5.** Colibra design fidelity in the final synthesized V<sub>HH</sub> library. **(a)** Observed CDR3 length distribution: 92% of the resulting library was observed in-frame. **(b)** and **(c)** Combination analysis of the amino acid incorporation efficiencies at 255 independent positions in the V<sub>HH</sub> library, as determined by high throughput sequencing, compared to expected values by design. **(b)** Per-amino acid log (observed/expected) incorporation fidelity, showing the median, 25th percentile, 75th percentile, and outliers. **(c)** Data for all amino acid incorporation frequencies compared with design displayed in linear-scale (Pearson's  $r = 0.963$ ).

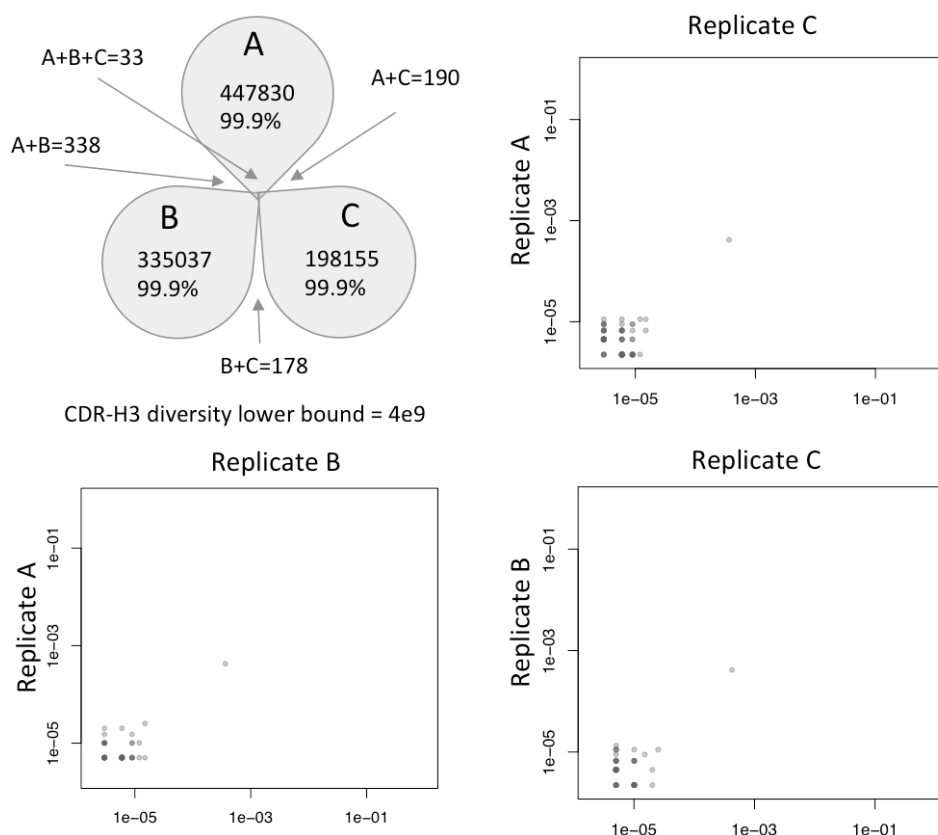
NGS data analysis of three sample replicates of the V<sub>HH</sub> CDR3s showed that the synthesis produced a very diverse library such that 99.9% of all sequences observed, even at this great depth, were unique (Figure 6). A control  $2e-2$  clone existed to calibrate enrichment during selection. A rare population of shared clones began to emerge in the shorter CDRs at  $1e-5$  (one in one hundred thousand sequences), although only 33 of such clones were reliably recovered in all three replicates. Using Fischer's capture

recapture, the diversity of the resulting synthesized material was calculated to be minimally 4 billion molecules, and potentially much higher. The results validate our prediction that no physical synthesis steps were restricting the library through any diversity bottlenecks that would adversely affect the resulting library quality.

**Table 2.** Quality assessment of V<sub>HH</sub> libraries. NGS sequence data was assessed to determine the percentage of CDR regions of the designed length.

Domain	Designed length %	Correct aa identity %	QC Pass <sup>a</sup> %
CDR1	93.92	99.7	100
CDR2	87.87	99.7	100
CDR3-Total	92.34	99.7	99.7

<sup>a</sup> Quality control (QC) pass refers to specified/expected (E) versus observed (O) percentage for the fabrication of the designed codons within the CDRs. The criteria were as follows: E0% = O<1%; E2-5% = O0.01-15%; E6-10% = O1-20%; E11-20% = O3-40%; E21-30% = O5-50%; E31-60% = O10-80%; E61-90% = O30-95%; E91-100% = O70-100%, where expected (E) values are rounded down to the nearest integer. CDR1 data collected from 3297977 reads; reads; CDR2 from 4764625 reads; CDR3 data collected from 2725295.



**Figure 6.** Synthesized molecular diversity of V<sub>HH</sub> CDR-H3. Three replicates were sequenced to depths of 447830, 335037 and 198155 reads, respectively. Over 99.9% of all reads in each replicate were unique and specific to the replicate in which they were generated, with only 33 clones observed across all three replicates at this depth of sequencing. A minimum CDR3 diversity of 4 billion unique CDR3s is calculated by capture-recapture. Plots show clone frequency overlap between replicates. A single 5e-3 enrichment control clone is observed in all libraries, with shared clones becoming evident with frequencies 1e-5 or below.

### 2.3. Discussion

In this study, we have extended the use of ProxiMAX beyond trimer addition to accommodate hexamer additions both for manual and automated mixing. We initially used the hexamers for manual addition in a semi-minimalist design for V<sub>L</sub> CDRs in which the most frequent amino acids in our database of V<sub>L</sub> sequences were simplified to the nearest 5%, with a minimal threshold of 5%. We observed a very close correlation of the observed to the expected (or designed) frequency for every position across the CDRs in this library, demonstrating the utility of this method.

We subsequently refined the technique for an automated platform which was used to construct the CDR3 loops of a camelid V<sub>HH</sub> library. These loops differ from the canonical structures typical of antibodies with paired V<sub>H</sub> and V<sub>L</sub> domains [15]—an adaptation that probably evolved to compensate for the lack of diversity which can result from pairing V<sub>H</sub> and V<sub>L</sub> domains. In general the CDR3 loops from V<sub>HH</sub> domains are longer than those of HCDR3s from human and mouse repertoires [3,4,15,18]. Fabrication of the V<sub>HH</sub> CDR3 loop lengths demonstrated a further enhancement over the trimer method in that loop lengths greater than 15 amino acids, which present difficulties in the assemblies with single trimer additions, were comfortably tackled with the hexamer components by building the CDR3 in two halves of up to 12 amino acids each and then assembling the completed loop. Hence, CDR3 regions up to 24 randomized amino acids were created. At this length, it is even more critical to be able to dictate functional amino acids as the theoretical diversity extends to 20<sup>24</sup> (1.7 × 10<sup>31</sup>) combinations, but through judicious choice of amino acids and their frequencies, this can be distilled to achieve greater sampling of the functional space. This concept has been previously exemplified, where aggregate frequencies of V<sub>H</sub> and V<sub>K</sub> CDR3 regions were incorporated into a Fab library design that resulted in 93% of clones displaying correctly folded heavy and light chains [6]. However, despite extreme examples of minimalist designs, incorporating amino acids which are over-represented in CDRs (primarily Tyr and Ser), being able to generate high affinity hits [19], severely restricted designs (Y/S) have been shown to lead to lower affinity binders than a fuller repertoire of amino acids [20]. Nevertheless, synthesis methods that allow programmable design of the CDR composition will permit better interrogation of the capacity of this site for antigen binding [5].

Whilst MAX codon libraries have been effectively utilized in our antibody engineering to produce libraries that are highly functional, to our knowledge the process described herein is the first report of a saturation mutagenesis method that accommodates a full set double codon hexamer blocks (20 × 20 codons) that can precisely randomize two neighboring positions in parallel in a single step. Clearly the work described herein is not for use at the standard laboratory bench, where the purchase of so many oligonucleotides (and the necessary automation to handle them) would be a prohibitive expense. Rather, we present the achievements of Colibra™ in order to allow prospective users to compare the results achieved using Colibra™ with previously-published achievements of SlonoMAX and TRIM technologies. It is for the accomplished antibody engineers to decide whether Colibra™ offers distinct advantages. However, we also believe that besides the performance enhancement in terms of speed and final purity, there are additional advantages that can be obtained through enzymatic ligation methods such as reducing problematic motifs in the randomized region. Specifically, there are problematic paired residue motifs that have a propensity to cause issues during manufacturing or long-term storage of antibodies which include NG, NS, NA (deamidation) and DG (isomerization) [4]. These biochemical liabilities can

be reduced within the initial library by excluding the corresponding hexamers; however certain pairs may exist at the ligation junctions. To eliminate these completely, depending upon design, a split-pot synthesis method could be employed. It is also conceivable that specific glycosylation signals can be removed through a similar approach so that asparagine can be safely encoded within the CDR.

Therefore, we have demonstrated that the ProxiMAX technique can be adapted to accommodate additions of paired randomized codons. The ProxiMAX method is particularly suited to this purpose as utilizes blunt-ended ligation of codon blocks so that complementary sticky-ends do not have to be created which then require precise pairing. Advantageously, using enzymatic ligations in preference to solid-phase chemical synthesis method, such as TRIM, provides the possibility of including most codons, which can be readily synthesized and used immediately within library fabrication. Some codons that encode a *MlyI* recognition site (GAGTC) are excluded from our process, therefore precluding the use of certain codon pairs such as GAG-TCN, although other synonymous codons can be used to encode the amino acids. Having a wide choice of available codons is relevant when specifically designing codon optimized libraries for production strains used in the manufacture of biologics. Consequently, the modified ProxiMAX method is particularly applicable to the discovery and development of biologics from hit identification to potentially optimizing the expression of sequences for manufacture.

### 3. Experimental Section

Oligonucleotides were purchased from Biosearch Technologies, CA and Integrated DNA Technologies (IDT). Enzymes were from New England Biolabs and ThermoFisher. Libraries were synthesized using a proprietary, automated version of ProxiMAX randomisation, from hairpin donor and acceptor sequences, essentially as described for the manual process [6], except that 400 hexameric donor sequences each containing two “MAX” codons were employed. The hexamer mixes were prepared by using a Tecan Freedom Evo liquid handler, fitted with disposable tips. Various parameters were modified on the robot to ensure accuracy. These included the development of a custom liquid class compatible with the physical properties of the diluted oligonucleotides, to optimize the aspiration and dispense steps of each hexamer in the mixing process.

Ligations of the hexameric mixes to the acceptor molecules and PCR amplification of the products were performed as previously described<sup>7</sup>, except that: (i) ligations were performed either individually (manual approach) or in pools (automated approach) using  $1 \times$  T4 Ligase buffer supplemented with PEG 4000; (ii) PCR amplifications were carried out in  $1 \times$  HF buffer, 200uM each dNTP, 0.3  $\mu$ M forward and reverse primer, 1U Phusion HSII Polymerase for every 50  $\mu$ L PCR reaction; 15 cycles max.

The product of the final cycle of MAXMAX codon additions was subjected to a final ligation to an ‘adapter’ oligo, to provide fixed framework regions for linkage to other gene segments by way of compatible overhangs generated by *BsaI* restriction enzyme digestion at sites engineered at precise locations in the adapter oligos. Typically, 50–100 pmol of material was digested using 20U *BsaI*-HF per  $\mu$ g DNA in 1X CutSmart buffer for at least 2 h at 37 °C. The product was then agarose or PAGE gel extracted, purified and quantified. Alongside, framework regions were prepared by PCR amplification of a sequence-verified gene fragment and similarly digested, gel extracted, purified and quantified.

Full-length gene libraries were assembled by ligating *BsaI*-treated mutagenized cassettes to framework regions that had compatible overhangs, using T4 DNA Ligase. Typically, 10 pmol of each

component were ligated and the ligation products then minimally amplified by PCR using Phusion HSII DNA Polymerase to generate the final library material. NGS sequencing was performed on an Illumina MiSeq sequencer using version 2 Miseq Reagent kits for amplicon analysis.

#### 4. Conclusions

Over recent years the performances of the first synthetic and semi-synthetic libraries have been critically assessed. Whilst these libraries have provided adequate hits to many pharmaceutically relevant targets, technologies now exist to define and interrogate more precisely functional library designs to generate higher affinity, more selective hits with better biophysical characteristics. The automated synthetic process described herein provides a method in which the most sophisticated schemes can be fabricated with exceptionally accuracy inasmuch as the required amino acids are encoded almost exclusively and compliance between designed and observed ratios of encoded amino acids is generally achieved within tight margins. In particular, our process incorporates hexamer codon additions that allow short, problematic, motifs to be reduced or excluded. The procedures are tightly monitored through next generation sequencing to provide precise quality control feedback on the fabrication method to ensure that the library closely matches the design. Thus, close correlation of design and manufacture in building sophisticated DNA libraries has immediate application in the discovery of new biological molecules.

#### Acknowledgments

We gratefully acknowledge O. Sanz for technical assistance, P. Mathonet for camelid repertoire NGS and analysis and B. Meineken for helpful discussions.

#### Author Contributions

LF designed the study, analyzed data and co-wrote the paper. MES contributed to the design of the study and performed experiments. CB, AS, SEC, NK, DE, SS and CB performed experiments. JG designed the  $V_{HH}$  libraries, analyzed data and co-wrote the paper. CGU designed the study,  $V_L$  libraries, analyzed data and co-wrote the paper. AVH analyzed data and co-wrote the paper.

#### Conflicts of Interest

The authors declare that the scFv and  $V_{HH}$  libraries described herein and the ProxiMAX process used to construct those libraries are commercial products marketed by Isogenica Ltd (UK). ProxiMAX is marketed as Colibra™. AVH further declares a financial interest inasmuch as she is a named inventor of the ProxiMAX process.

#### References

1. Ponsel, D.; Neugebauer, J.; Ladetzki-Baehs, K.; Tissot, K. High affinity, developability and functional size: The holy grail of combinatorial antibody library generation. *Molecules* **2011**, *16*, 3675–3700.
2. Hoogenboom, H.R. Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **2005**, *23*, 1105–1116.

3. Finlay, W.J.J.; Almagro, J.C. Natural and man-made V-gene repertoires for antibody discovery. *Front. Immunol.* **2012**, *3*, eArticle 342.
4. Strohl, W.R.; Strohl, L.M. *Therapeutic Antibody Engineering: Current and Future Advances Driving the Strongest Growth Area in the Pharmaceutical Industry*, 1st ed.; Woodhead Publishing: Amsterdam, Netherlands, 2012.
5. Mahon, C.M.; Lambert, M.A.; Glanville, J.; Wade, J.M.; Fennell, B.J.; Krebs, M.R.; Armellino, D.; Yang, S.; Liu, X.; O'Sullivan, C.M.; *et al.* Comprehensive Interrogation of a Minimalist Synthetic CDR-H3 Library and Its Ability to Generate Antibodies with Therapeutic Potential. *J. Mol. Biol.* **2013**, *425*, 1712–1730.
6. Zhai, W.; Glanville, J.; Fuhrmann, M.; Mei, L.; Ni, I.; Sundar, P.D.; Van Blarcom, T.; Abdiche, Y.; Lindquist, K.; Strohner, R.; *et al.* Synthetic antibodies designed on natural sequence landscapes. *J. Mol. Biol.* **2011**, *412*, 55–71.
7. Ashraf, M.; Frigotto, L.; Smith, M.E.; Patel, S.; Hughes, M.D.; Poole, A.J.; Hebaishi, H.R.M.; Ullman, C.G.; Hine, A.V. ProxiMAX randomization: A new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochem. Soc. Trans.* **2013**, *41*, 1189–1194.
8. Kille, S.; Acevedo-Rocha, C.G.; Parra, L.P.; Zhang, Z.-G.; Opperman, D.J.; Reetz, M.T.; Acevedo, J.P. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol.* **2013**, *2*, 83–92.
9. Tang, L.; Wang, X.; Ru, B.; Sun, H.; Huang, J.; Gao, H. MDC-Analyzer: a novel degenerate primer design tool for the construction of intelligent mutagenesis libraries with contiguous sites. *Biotechniques* **2014**, *56*, 301–302, 304, 306–308.
10. Glanville, J.; Zhai, W.; Berka, J.; Telman, D.; Huerta, G.; Mehta, G.R.; Ni, I.; Mei, L.; Sundar, P.D.; Day, G.M.; *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 20216–20221.
11. Van den Brulle, J.; Fischer, M.; Langmann, T.; Horn, G.; Waldmann, T.; Arnold, S.; Fuhrmann, M.; Schatz, O.; O'Connell, T.; O'Connell, D.; *et al.* A novel solid phase technology for high-throughput gene synthesis. *Biotechniques* **2008**, *45*, 340–343.
12. Ashraf, M.; Hughes, M.D.; Hine, A.V. Oligonucleotide library encoding randomised peptides. Patents EP 1907548, 2011; EP 2236612, 2012 and US 8357638 B2 2013.
13. Tiller, T.; Schuster, I.; Deppe, D.; Siegers, K.; Strohner, R.; Herrmann, T.; Berenguer, M.; Poujol, D.; Stehle, J.; Stark, Y.; *et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* **2013**, *5*, 445–470.
14. Chumsae, C.; Gaza-Bulseco, G.; Sun, J.; Liu, H. Comparison of methionine oxidation in thermal stability and chemically stressed samples of a fully human monoclonal antibody. *J. Chromatogr. B* **2007**, *850*, 285–294.
15. Muyldermans, S. Single domain camel antibodies: Current status. *Rev. Mol. Biotech.* **2001**, *74*, 277–302.
16. Wang, F.; Ekiert, D.C.; Ahmad, I.; Yu, W.; Zhang, Y.; Bazirgan, O.; Torkamani, A.; Raudsepp, T.; Mwangi, W.; Criscitiello, M.F.; *et al.* Reshaping antibody diversity. *Cell* **2013**, *153*, 1379–1393.
17. Johnson, G.; Wu, T.T. Kabat Database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* **2000**, *28*, 214–218.

18. Muyldermans, S.; Baral, T.N.; Cortez Retamozzo, V.; De Baetselier, P.; De Genst, E.; Kinne, J.; Leonhardt, H.; Magez, S.; Nguyen, V.K.; Revets, H.; *et al.* Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* **2009**, *128*, 178–183.
19. Fellouse, F.A.; Esaki, K.; Birtalan, S.; Raptis, D.; Cancasci, V.J.; Koide, A.; Jhurani, P.; Vasser, M.; Wiesmann, C.; Kossiakoff, A.A.; *et al.* High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J. Mol. Biol.* **2007**, *373*, 924–940.
20. Hackel, B.J.; Wittrup, K.D. The full amino acid repertoire is superior to serine/tyrosine for selection of high affinity immunoglobulin G binders from the fibronectin scaffold. *Protein Eng. Des. Sel.* **2010**, *23*, 211–219.

© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).