Construction and maintenance of randomized retroviral expression libraries for transmembrane protein engineering

Sara A. Marlatt 1, Yong Kong 2,3, Tobin J. Cammett 1,6, Gregory Korbel 1,7, James P. Noonan 1,5 and Daniel DiMaio 1,2,4,5,8

1Department of Genetics, Yale School of Medicine, PO Box 208005, New Haven, CT 06520-8005, USA, 2Department of Molecular Biophysics and Biochemistry, Yale School of Medicine, 300 George Street, 8th Floor, New Haven, CT 06511-6624, USA, 3Keck Biotechnology Resource Laboratory, Yale University, 300 George Street, Box 201, New Haven, CT 06511, USA, 4Department of Therapeutic Radiology, Yale School of Medicine, PO Box 208040, New Haven, CT 06520-8040, USA, 5Yale Comprehensive Cancer Center, PO Box 208028, New Haven, CT 06520-8028, USA, 6Present address: The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

5Present address: Alexion Pharmaceuticals, 352 Knotter Drive, Cheshire, CT 06410, USA

8To whom correspondence should be addressed.
E-mail: daniel.dimaio@yale.edu

Received October 29, 2010; revised October 29, 2010; accepted November 14, 2010

Edited by Arthur Horwich

Genetic selection from libraries expressing proteins with randomized amino acid segments is a powerful approach to identify proteins with novel biological activities. Here, we assessed the utility of deep DNA sequencing to characterize the composition, diversity, size and stability of such randomized libraries. We used 454 pyrosequencing to sequence a retroviral library expressing small proteins with randomized transmembrane domains. Despite the potential for unintended random mutagenesis during its construction, the overall hydrophobic composition and diversity of the proteins encoded by the sequenced library conformed well to its design. In addition, our sequencing results allowed us to calculate a more accurate estimate of the number of different proteins encoded by the library and suggested that the traditional methods for estimating the size of randomized libraries may overestimate their true size. Our results further demonstrated that no significant genetic bottlenecks exist in the methods used to express complex retrovirus libraries in mammalian cells and recover library sequences from these cells. These findings suggest that deep sequencing can be used to determine the quality and content of other libraries with randomized segments and to follow individual sequences during selection.

Keywords: bovine papillomavirus/E5 protein/randomization/sequencing/traptamers

Introduction

A major goal of protein engineering is to construct proteins or polypeptide segments that interact with native cellular or viral proteins and modulate their activity. This is a formidable task because protein–protein interaction surfaces are usually generated by complex three-dimensional folding patterns, often involving amino acids that are not contiguous along the polypeptide chain. The identification of polypeptides with biological activity is further complicated by the fact that most short protein segments do not adopt a stable conformation (Norman et al., 1999; Peelle et al., 2001). A popular approach to circumvent these limitations and identify proteins or peptide segments with novel activities is to randomize a short stretch of amino acids and then analyze the product. Using this approach, we assessed the utility of deep DNA sequencing to characterize the composition, diversity, size and stability of such randomized libraries. We used 454 pyrosequencing to sequence a retroviral library expressing small proteins with randomized hydrophobic amino acid sequences. Most hydrophobic peptide segments of a sufficient length (~20 amino acids) are predicted to adopt a stable α-helical configuration that can span a lipid bilayer due to the formation of intrachain hydrogen bonds that sequester the main-chain amino and carbonyl groups away from the lipids (Engelman et al., 1986). Thus, in contrast to the situation for soluble, globular protein scaffolds, where most random sequences are expected to be unstructured, this approach is likely to be permissive for a large majority of hydrophobic amino acid sequences. We designate proteins containing these random TM aptamers as ‘traptamers’. We constructed traptamer expression libraries based on the bovine papillomavirus (BPV) E5 protein, a small TM protein with potent oncogenic activity (Schlegel et al., 1986). Only 44-amino acids long, the E5 protein is essentially an isolated TM domain that spans intracellular membranes of transformed cells as a disulfide-linked homodimer (Schlegel et al., 1986; Burkhart et al., 1989; Surti et al., 1998). The E5 protein is thought to interact directly with the TM domain of two molecules of the platelet-derived growth factor (PDGF) β receptor in a highly specific fashion, resulting in dimerization of the receptor, activation of its intrinsic tyrosine kinase activity, and mitogenic signaling (Petti et al., 1991; Petti and DiMaio, 1992; Nilson et al., 1995; Lai et al., 1998; Nappi and Petti, 2002). Our analysis of the E5 protein
suggested that it might be possible to develop small, artificial TM proteins that could specifically interact with and modulate the activity of other cellular or viral TM proteins (Freeman-Cook et al., 2004). This approach could have wide applicability because genome sequence analysis suggests that up to 30% of cellular proteins and many viral proteins contain membrane-spanning domains (Lehnert et al., 2004).

To utilize the E5 protein as a scaffold to display diverse hydrophobic sequences to potential TM target proteins, we constructed libraries encoding a large number of small TM proteins in which the central hydrophobic segment of the E5 protein was replaced with randomized hydrophobic amino acids. By expressing these libraries from retroviral vectors in mammalian cells and using genetic techniques to select active library members, we readily isolated small proteins with diverse TM sequences that specifically bound and activated the PDGF β receptor, the natural target of the BPV E5 protein (Freeman-Cook et al., 2004, 2005; Freeman-Cook and DiMaio, 2005; Ptacek et al., 2007). These results validated the use of the E5 protein as a scaffold from which to express a wide array of TM domains and confirmed that diverse TM sequences were able to recognize and activate the PDGF β receptor.

In order to extend this approach to TM targets other than the PDGF β receptor, we redesigned the randomized segment of E5 so that it was longer than in our previous libraries and more closely matched the amino acid composition of naturally occurring TM domains. The resulting library, named TJC-5, was designed to contain a 19-amino acid, randomized TM segment in which ~82% of the amino acids were hydrophobic. The TJC-5 library was screened for traptamers that activated the human erythropoietin receptor (hEPOR). We isolated a 44-amino-acid protein (TC2-3) that was unrelated to erythropoietin but specifically activated the hEPOR and drove erythroid differentiation of primary human hematopoietic progenitor cells (Cammett et al., 2010).

To optimize the construction and use of protein expression libraries with randomized segments, it is important to verify that the libraries conform to their design, establish a rigorous estimate of the number of different TM proteins encoded by these libraries and determine whether this genetic complexity is maintained during retrovirus production, infection and recovery from mammalian cells. Here, we used deep DNA sequencing to determine the composition, complexity, diversity and stability of the TJC-5 library. Our sequencing results showed that the E5 protein is a suitable scaffold for construction and expression of complex, highly diverse TM protein libraries that are stably maintained during retroviral packaging and infection of mammalian cells.

**Materials and methods**

**Cells and viruses**

Human embryonic kidney (HEK) 293T cells were maintained in the Dulbecco’s Modified Eagle’s Medium (DMEM) supplemented with 10% fetal bovine serum (FBS, Gemini Bioproducts), 4 mM l-glutamine, 20 mM HEPES (pH 7.3) and 1 × penicillin/streptomycin (Gibco; DMEM-10). Interleukin (IL)-3-dependent, murine BaF3 cells were maintained in RPMI-1640 medium supplemented with 5% WEHI-conditioned medium (as a source of IL-3), 10% heat-inactivated FBS, 2 mM l-glutamine, 0.06 mM β-mercaptoethanol and 1 × penicillin/streptomycin (RPMI/IL-3). All cells were cultured in 5% CO₂ at 37°C.

Recombinant retrovirus particles were prepared by calcium phosphate-mediated co-transfection of HEK 293T cells with pan tropic pSVV-G (Clontech), pCL-Eco (Navaux et al., 1996) and TJC-5 library plasmid DNA. Transfected cells were cultured in OptiMEM (Gibco) for 48 h at 37°C, and then infectious virus particles were recovered, filtered using a 0.45-μM syringe filter (Millipore) and concentrated using Centricon Ultracel PL-30 tubes (Millipore).

**Plasmids and cloning**

A modified, full-length BPV E5 gene was designed to include a canonical Kozak sequence (accatgCCG) by replacing the nucleotides surrounding the start codon in the wild-type gene (actatgCCGA; Kozak, 1986). This generated a proline to alanine substitution at E5 codon 2. In addition, AvrII and BstBI restriction sites were introduced into the E5 sequence by silent mutations in codons 10 and 35–36, respectively. This was accomplished by annealing oligonucleotides with complementary 3’ ends and extending them by PCR. The double-stranded product was purified (Qiagen), digested with BamHI and Xhol (New England Biolabs) and ligated into digested pRV-HygR, a retroviral vector that encodes a hygromycin resistance gene (Riese and DiMaio, 1995; Freeman-Cook et al., 2005). The AvrII restriction site in the pRV-HygR backbone was removed by QuickChange site-directed mutagenesis. This final construct was denoted pRV-E5/p2A (see Supplementary Table S1 for oligonucleotide sequences).

To facilitate efficient cloning of library sequences into the pRV-E5/p2A backbone, a 1.3-kb sequence from pCMMP-IRES-GFP (a gift from Bill Sugden, University of Wisconsin) was cloned into pRV-E5/p2A using SpeI and BamHI restriction sites to replace codons 14–45 of the E5/p2A sequence with the IRES-GFP sequence. This new retroviral plasmid was named pT2H-F13 and encodes a bicistronic transcript consisting of the first 13 residues of the E5/p2A gene followed by the IRES-GFP sequence.

**Retroviral library construction**

To design a library that was primarily hydrophobic but also included hydrophilic and small amino acids, we first calculated the expected frequency of each amino acid within the randomized segment for any proposed ratio of nucleotides (the spreadsheet used for this calculation can be found at www.med.yale.edu/genetics/dimaio/index.html). By assessing the coding potential of various ratios of the four nucleotides, we empirically identified a mixture that encoded amino acids at frequencies similar to those found in natural TM domains (Larsson et al., 2002). In this mixture, the ratio of A:C:G:T was 1:1:1:0.5 at the first position in each codon and 0:1:0.25:0.1:1 at the second position. At the third position, A and T were excluded, and C and G were present in a 1:0.1 ratio. To construct the library, a degenerate oligonucleotide was synthesized (Yale Program for Critical Technologies in Molecular Medicine) which encoded residues 9–39 of the E5/p2A sequence, with residues 12–30 being randomized with the nucleotide ratios specified above. In addition, at residue 33, position 1 was an equimolar mixture of A, C and G, position 2 was restricted to A, and position 3 was an
equimolar mixture of C and G. These nucleotide mixtures allowed only strongly hydrophilic amino acids (aspartic acid, glutamic acid, histidine, glutamine, asparagine and lysine) at this position.

An invariant oligonucleotide complementary to residues 34–44 of the E5/P2A sequence (including a stop codon and restriction sites for cloning) was annealed to the 3' end of the degenerate oligonucleotide and extended (Fig. 1). The double-stranded product was purified (Qiagen) and amplified by PCR for 20 cycles using primers complementary to the invariant sequences at each end (see Supplementary Table S1 for oligonucleotide sequences). The full-length amplified product was purified, digested with AvrII and BamHI and ligated into digested pT2H-F13 to reconstitute an intact open reading frame. The ligation mixture was purified (Qiagen) and used to transform *Escherichia coli* strain DH10β (Invitrogen) by electroporation. Lawns of ampicillin-resistant bacteria, representing ~1.2 million independent transformants, were pooled and processed by the Maxiprep protocol (Qiagen) to purify the plasmid library DNA, which was named TJC-5. Additional details regarding PCR and cloning are available from the authors upon request.

**Infection and recovery from mammalian cells**

BaF3 cells expressing the hEPOR (BaF3-hEPOR) were described previously (Cammett *et al.*, 2010). To introduce the TJC-5 library into clonal BaF3-hEPOR cells, 500,000 cells/well were plated in a 12-well plate in 500 μl of RPMI/IL-3. Two wells were infected with 500 μl of concentrated TJC-5 library retrovirus stocks and supplemented with polybrene at a final concentration of 4 μg/ml (Sigma). Infections were incubated at 37°C for 4 h and then transferred to separate T25 flasks (Falcon) containing 9 ml of RPMI/IL-3 and 9 μl of polybrene. IL-3 was included to eliminate selective pressure for the expression of an active TM protein. Infected cells were split 1:2 on days 2, 4 and 6 post-infection by transferring 5 ml of resuspended cells into 5 ml of fresh RPMI/IL-3 medium containing 2 mg/ml hygromycin B (Invitrogen).

**454 sequencing of the TJC-5 library**

For sequencing of the plasmid library, genomic DNA (gDNA) was first prepared from BaF3-hEPOR cells using the DNeasy Blood and Tissue Kit (Qiagen). Ten micrograms of Sspl-linearized TJC-5 plasmid DNA was then added to this gDNA sample to more closely mimic the reaction conditions of the post-infection recovery process described below. Five hundred nanograms of this DNA mixture was used as a template for PCR (Expand Long Template, Roche), in which library sequences were amplified for 30 cycles using primers specific for the 5' and the 3' invariant regions of the E5 sequence retained in the library. A single PCR product was observed (approximate length of 164 bp). Two micrograms of purified PCR product from two independent sets of reactions was submitted separately to the Yale University Center for Genomics and Proteomics for sequencing using a Roche/454 GS FLX sequencing analyzer and long read 70 sequencing preparation. For sequencing of the recovered library, gDNA was isolated separately from ~3.5 × 10⁸ hygromycin-resistant BaF3-hEPOR cells for each of the two independent pools of library-infected cells. PCR recovery was carried out separately for each recovered library sample, and 2 μg of each purified PCR product was submitted for sequencing. One-quarter of the sequencing plate was used for each sample. The sequences obtained were deposited to the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra, accession number SRA022064).

**Sequence analyses**

**Sequence trimming.** The invariant sequence flanking the randomized TM segment was analyzed to assess the quality of each sequencing read. First, an algorithm was designed (BvTrim; http://graphics.med.yale.edu/trim/) to search the 5' invariant region for the 32 nucleotides directly upstream of the randomized segment. Sequences with more than four mutations (insertions, deletions or substitutions) within the 5' invariant region were removed from the sequence set. The same parameters were then used to analyze the 3' invariant region for the 32 nucleotides directly downstream of the randomized segment, and any sequences with more than four mutations in the 3' invariant region were also removed from the sequence set.

**Homopolymer collapsing, grouping and signature re-expansion.** Signature sequences were generated by collapsing any run (two or more) of the same nucleotide into a single representative nucleotide at a given position. These simplified sequences were then compared with one another, and identical sequences were grouped together. For each group, the original, full-length sequences were then sorted according to the length of the randomized segment, and the most abundant sequence in each group was chosen as a representative sequence. If two sequences were equally abundant within a group, one was chosen randomly to be the representative sequence. Those representative sequences with insert lengths of 66 bp were analyzed further as described.
Fig. 2 Amino acid sequence of wild-type BPV E5, the TJC-5 library and an active clone. The 44-amino-acid sequence of the BPV E5 protein is shown on the first line. The wild-type E5 sequence and the invariant residues retained in the TJC-5 library (second line) and TC2-3 (third line) are shown in bold. X’s represent residues randomized in the TJC-5 library design. TC2-3 was selected from the TJC-5 library and activates the hEPOR (Cammett et al., 2010).

![Amino acid sequence of wild-type BPV E5, the TJC-5 library and an active clone.](https://example.com/fig2.png)

Fig. 3 Amino acid composition of the randomized segment of the TJC-5 plasmid library. The frequency of each amino acid is given for naturally occurring TM domains (black), and for the 19 consecutive randomized codons in the TJC-5 plasmid library. The predicted composition of the randomized segment is shown in grey bars and the observed composition, based on the sequencing data, is shown in white bars. Asterisk represents the stop codons.

![Amino acid composition of the randomized segment of the TJC-5 plasmid library.](https://example.com/fig3.png)

Pairwise diversity comparisons. For the diversity analysis, the randomized positions (60 nucleotides corresponding to codons 12–30 and 33) within a given, full-length sequence were compared with the randomized positions of every other full-length sequence in a pairwise fashion. For each pair, a distance score was generated, defined as the total number of positions at which the corresponding nucleotides differed between the sequences.

Defined library simulation. For the defined library simulation, samples of a given size, ranging from 1 to 75% of the total size of the defined library (93,245), were chosen randomly from the defined library. The chosen sequences were then compared with a second, independently chosen sample set of the same size, and the number of sequences in common between the two samples was determined. Two hundred paired sample sets were compared for each sample size, and the number of sequences in common for each comparison was determined and used to calculate library size according to the capture–recapture equation.

Simulation to determine overlap between sequence sets. The overlap between each pair of samples was determined by comparing 36,031 non-redundant, full-length sequences from each sample being compared. The smallest sample, P1, consists of 36,031 such sequences. For samples P2, R1 and R2, 36,031 sequences were selected at random from the larger number of sequences. Two hundred paired sequence sets were generated for each sample comparison. The number of sequences in common between sequence sets selected from different samples was then determined.

Results and discussion

Design of the library

The TJC-5 library was designed to express a large number of different, 44-amino-acid TM proteins (traptamers), each of which contains a putative TM domain comprised of a random sequence of primarily hydrophobic amino acids (Fig. 2). We constructed this library using an oligonucleotide template in which a randomized segment was synthesized using non-equimolar mixtures of each of the four nucleotides (see the ‘Materials and methods’ section), such that most of the resulting codons encoded hydrophobic amino acids. To increase the random surface presented to targets and to minimize any potential preference for the PDGF β receptor by the remaining E5 sequences, we increased the length of the randomized segment compared with our previous libraries to nineteen consecutive codons, corresponding to amino acids 12–30 of the wild-type E5 sequence. Although the randomized segment is shorter than a typical TM domain, it is flanked at its amino terminal end by eight fixed hydrophobic amino acids from the wild-type E5 protein. Because pairs of aromatic amino acids are frequently found at the ends of TM domains, where they are thought to help anchor the TM protein near the polar head groups of the lipid bilayer (White and Wimley, 1999), the randomized segment was followed by two wild-type codons (codons 31 and 32) encoding the aromatic amino acids tyrosine and tryptophan. Codon 33, which encodes an essential aspartic acid in the E5 protein (Klein et al., 1999), was randomized to encode only strongly hydrophilic amino acids to allow for potential interactions with charged residues in the juxtamembrane region of target proteins. The proteins encoded by this library retained the 11 N-terminal and 11 C-terminal amino acids from the E5 protein, which are thought to protrude from the membrane bilayer. The invariant C-terminus includes the cysteines involved in covalent dimer formation by the wild-type E5 protein, so some of the encoded proteins may be able to form disulfide-linked dimers, which might be advantageous by increasing the surface area and the number of amino acids available to bind to a potential target. Indeed, homodimerization is essential for the biological activity of TC2-3, the novel hEPOR activator recovered from this library (Cammett et al., 2010).

We designed the randomized segment to have an amino acid composition that was similar to that of naturally occurring TM domains, in which ~80% of amino acids are hydrophobic (Larsson et al., 2002; Liu et al., 2002), so that the expressed proteins would be sufficiently diverse to recognize a wide variety of targets. As shown in Fig. 3, the strongly hydrophobic amino acids leucine, valine, isoleucine and phenylalanine were predicted to be the most common amino acids within the randomized segment. Alanine, serine and threonine were also relatively abundant, as they are in natural TM domains, while all other amino acids were encoded at low frequency. In contrast, the TM domain of the E5 protein, which recognizes only the PDGF β receptor, contains a single hydrophilic amino acid. Although TM proteins shorter than the full-length E5 protein can be biologically active
(Talbert-Slagle et al., 2009), the predicted low frequency of stop codons within the randomized segment (0.1%) should maximize the number of full-length library proteins.

Library construction, sequencing and quality control

To construct the TJC-5 library, we amplified the oligonucleotide containing the randomized segment and cloned it into a retroviral expression vector (Fig. 1). Plasmid DNA was isolated from ~1.2 million pooled bacterial transformants and designated TJC-5. The library coding sequences were amplified from the plasmid DNA and subjected to 454 pyrosequencing. The small size of the library amplicons made it possible to sequence each randomized segment (66 nucleotides, representing codons 12–33) in a single run, thus eliminating the need for contig assembly. This was essential to our analyses, because different randomized clones could share a significant block of sequence identity yet be distinct, making other high-throughput sequencing approaches with shorter read lengths or higher error rates unsuitable for this application.

Two independent PCRs were conducted and sequenced separately for the TJC-5 library (P1 and P2), yielding 202 227 sequences in total. Preliminary analysis of the sequencing results indicated that the quality and the composition of P1 and P2 were similar, so the two samples were combined in the analysis that follows, unless indicated otherwise. To reduce the influence of sample preparation and sequencing artifacts in subsequent analyses, several quality-control measures were employed (Fig. 4). First, sequences with more than four mutations in either of the invariant regions flanking the randomized segment (~10% of all sequences) were eliminated from the data set (Fig. 4, B2). The great majority of the sequences in the remaining 181 338 sequences were present only once. Next, we determined how many different sequences existed in these 181 338 sequences. Because 454-based sequencing has difficulty in correctly calling runs of identical bases, which can result in insertions or deletions in the sequencing reads (Margulies et al., 2005; Huse et al., 2007), any homopolymers in a given sequence were collapsed to a single nucleotide to generate a signature sequence (sequences lacking homopolymeric runs were also defined as signature sequences), and identical signatures were grouped. The most abundant sequence from each group was chosen as the representative sequence to generate a non-redundant data set of 120 333 different sequences (Fig. 4, C1). Although this approach may undercount the number of different sequences in the library, it allowed us to establish a stringent collection of different sequences and simplified the computations required for subsequent analyses.

Next, using the most abundant original sequence corresponding to each signature sequence, the length of the randomized segment in these non-redundant sequences was analyzed. Based on the design of the library, the 5’ and the 3’ invariant regions should flank 66 nucleotides, corresponding to the 57 positions encoding randomized codons 12–30, six wild-type nucleotides comprising codons 31 and 32, and the randomized codon 33. Of the 120 333 non-redundant sequences with the correct invariant ends, 93 245 sequences (>75%) had randomized segments of the expected size (Fig. 4, D1). The composition and the diversity of the plasmid library were determined based on the sequences retained in set D1.

Composition of the TJC-5 library

Using this high-quality, non-redundant sequence set (Fig. 4, D1), the nucleotide composition at each randomized position was determined. In general, the 19 consecutive randomized codons followed the design of the library (Fig. 5, plasmid library). The first position of each codon was composed of all four nucleotides, with T being the least abundant. The second position was mostly T, driving the overall hydrophobic amino acid composition of the randomized segment, and the third position was primarily C. The first two randomized codons deviated from the expected nucleotide ratios, presumably as a consequence of the oligonucleotide synthesis procedure during the transition from the addition of a single, specified nucleotide to the addition of a mixture of nucleotides. In addition, G was over-represented at the expense of C and A, most notably at position 1. The basis for this over-representation may be due to the higher coupling rate of phosphoramidites containing G compared with those containing C and A (Hall et al., 2009). The nucleotide composition of codon 33 also displayed the expected randomization pattern: almost exclusively A, C and G at the first position, A at the second position, and equimolar C and G at the third position (Table I).

As expected from the nucleotide analysis, in silico translation of the sequences yielded TM domains with a primarily hydrophobic amino acid composition, similar to that predicted by the library design (Fig. 3). All amino acids were represented, but nearly 85% of the encoded residues within the randomized segment were hydrophobic and 15% were...
hydrophilic, close to the predicted frequencies of 82 and 18%, respectively. Due to the increased frequency of G at the first position of each codon, valine (encoded by GTX) was the most prevalent amino acid, rather than the predicted leucine. At codon 33, asparagine, histidine, lysine, aspartic acid and glutamine were relatively abundant, while glutamic acid was present in only 2% of the sequences because of a paucity of the GAG codon (data not shown). Thus, this large, direct sampling of the TJC-5 library confirmed that the nucleotide and amino acid composition of the randomized segment, in general, conformed well to its design and that biased codon usage can be effectively used to generate

![Fig. 5 Nucleotide composition of the randomized segment.](image_url)

**Table 1.** Nucleotide composition and mutation frequency of codon 33<sup>a</sup>

<table>
<thead>
<tr>
<th>Position</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>33.3</td>
<td>33.3</td>
<td>33.3</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>50.0</td>
<td>50.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Plasmid</td>
<td>29.4</td>
<td>42.1</td>
<td>28.1</td>
<td>0.4</td>
<td>95.2</td>
<td>0.4</td>
<td>0.3</td>
<td>4.1</td>
<td>0.1</td>
<td>53.5</td>
<td>46.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Recovered</td>
<td>28.8</td>
<td>45.2</td>
<td>25.9</td>
<td>0.1</td>
<td>99.3</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.2</td>
<td>53.8</td>
<td>46.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The fraction of each nucleotide (as a percent of the total) is given for each position in codon 33 for the predicted (top), plasmid (middle) and recovered (bottom) libraries.

<sup>a</sup>Numbers may not add up to 100 due to rounding.
Occurrence of mutations in the library

In addition to the deliberate randomization described above, it was possible that mutations were introduced during oligonucleotide synthesis, PCR amplification and other steps in library construction. In addition, sequencing errors cannot be distinguished from true mutations, and although they would not affect the composition of the actual library, they nonetheless influence the analysis of the sequencing results.

Most insertions and deletions in the randomized segment would generate frameshift mutations which would disrupt the encoded protein and thereby reduce the number of functional proteins encoded by the library. To identify potential insertions and deletions, the length of each randomized segment was determined for non-redundant library sequences with high-quality invariant ends (Fig. 4, C1). As shown in Fig. 6, 77% of the sequences had the expected 66 nucleotide stretch between the invariant ends, another 12% were one nucleotide longer or shorter than the expected length and the remaining sequences had larger insertions and deletions. As noted above, 454 sequencing technology has difficulty in correctly sequencing homopolymers; therefore, the actual occurrence of frameshift mutations is probably rarer than this analysis suggests. In addition, it is possible that a frameshift mutation in the 5′ end of the randomized segment might be rectified by a second, compensatory downstream mutation that restored the correct sequence length and reading frame. However, our data suggest that such events were infrequent because if this occurred, the nucleotide and amino acid composition between the two mutations would deviate from the design.

We estimated the frequency of substitution mutations by examining the positions where the design of the library excluded one or more nucleotides. We determined the frequency of excluded nucleotides at these positions for the non-redundant sequences with randomized segments of the correct length. The library design allowed no A’s or T’s at the third position of codons 12–30. On average, at each codon in the randomized segment, an A or T was present at the third position 0.72% of the time. When codons 12 and 13 are not considered due to their more dramatic deviation from the design, the presence of a substitution mutation at the remaining 17 codons is reduced to just 0.21% per codon. In fact, 93% of the sequences had no A’s or T’s in the third position of any of the codons within the central, randomized segment. Additionally, because of the degeneracy of the genetic code, most third-position mutations will not affect the amino acid sequence of the proteins in the library, and we note that amino acid substitutions would in general be beneficial by causing additional diversification of the library. At codon 33, T was excluded from position 1, C, G and T were not allowed at position 2, and A and T were excluded from position 3 (Table I). At this codon, the incidence of excluded nucleotides was 0.40 and 0.15% for positions 1 and 3, respectively. At position 2, where the design allowed only A, the other three nucleotides (overwhelmingly T) were present in 4.8% of the sequences, resulting in the insertion of a hydrophobic amino acid at codon 33 ~4% of the time. For the invariant codons 31 and 32, non-wild-type amino acids were present 0.59 and 0.78% of the time, respectively (data not shown). Thus, <15% of the clones in the library accumulated substitution mutations during construction and assembly, most of which were likely silent.

Diversity of the TJC-5 plasmid library

It is possible that during the construction of the library, a subset of the sequences were amplified and subsequently underwent limited diversification by PCR, so that ultimately the library represented only a relatively small number of founding sequences, each giving rise to a cluster of closely related variants. To assess how different the sequences were from one another, we conducted pairwise comparisons between all 93 245 non-redundant, full-length TM nucleotide sequences (Fig. 4, D1). Each pair of TM sequences was given a distance score, defined as the total number of randomized positions where the two sequences differed from one another. A high distance score indicates that two sequences are very different from one another, whereas a low distance score indicates that the sequences share many nucleotides in common and likely descended from the same original sequence. Based on the design of the library, taking into account the prescribed abundance of each nucleotide and assuming no simplification and re-diversification of the library, a normal distribution was generated that described the probability that two randomly chosen sequences would have a given distance score. The most likely distance score for any pairwise comparison was ~27 (Fig. 7), meaning that if any two sequences were compared at random, there would be the highest probability that they would differ at 27 of the 60 randomized nucleotide positions. This number was not higher in part because the design of the library dictated the insertion of a C at the third position of codons 12–30 nearly 90% of the time. The peak of the observed probability distribution of the pairwise comparisons of the actual sequences was centered at a distance score of 28, nearly overlapping with the predicted distribution (Fig. 7). The presence of a small population of sequences with higher-than-expected distance scores (35–45) may be the result of the increased incidence of T in the third position of codon 33. This analysis

Fig. 6 Insert length distribution of the randomized segment. The distribution of lengths of the randomized insert is given for the plasmid library. The design of the library dictates that the randomized insert should be 66 bp, corresponding to codons 12–33. Lengths between 60 and 72 nucleotides are shown. Inserts smaller or larger than this length, in aggregate, comprised ~5% of sequences.
indicates that the TJC-5 library is in fact highly diverse, as designed, and not comprised of clusters of closely related sequences.

**Complexity of the TJC-5 plasmid library**

Based on the design of the TJC-5 library, it could theoretically express more than $10^{25}$ different, full-length proteins. However, the actual number of sequences in the TJC-5 plasmid library was far smaller for several reasons, including the amount of degenerate oligonucleotide originally amplified, the efficiency of the amplification and cloning steps, the amount of DNA used to transform *E.coli* and the transformation efficiency. The size of randomized protein expression libraries is traditionally estimated by counting the number of bacterial colonies or phage plaques that are pooled to generate the library (Golden et al., 1995; Kitamura et al., 1995; Hale et al., 2004). In the case of the TJC-5 plasmid library, ~1.2 million bacterial transformants were pooled. While this defines the maximum potential size of the library, it is not a valid estimate of the number of different proteins expressed by the library because it does not account for the existence of multiple copies of the same sequence or the presence of sequences unable to encode the desired proteins.

We analyzed our sequencing results to calculate a more accurate estimate of the number of different full-length sequences actually present in the TJC-5 plasmid library. Because individual transformants may have given rise to multiple copies of the same sequence due to preferential amplification during the preparation of samples for sequencing, we used a capture–recapture sampling technique employed by ecologists, epidemiologists and population geneticists to estimate library size (Seber, 2000). In this technique, a population is sampled independently on two occasions, and the number of times that the same individual is observed in the two samples is used to calculate the total size of the population using a probability equation $P(A \cap B) = P(A) \times P(B)$, $P(A)$ and $P(B)$ are the probabilities of selecting two independent samples, $A$ and $B$, from the population, and $P(A \cap B)$ is defined as the probability of the intersection, or common selections, between the two samples. The probability of selecting a given sample is defined as the number of selections made divided by the total size of the population, $T$. Thus, the equation can be re-written as $(A \cap B)/T = A/T \times B/T$, which simplifies to $T = A \times B/(A \cap B)$. According to this equation, it will become less likely that the same individual (or sequence) will appear in two independent samples as the total size of the population increases. In order to utilize this equation to estimate the size of the TJC-5 plasmid library, we separately processed the sequences from our two sequencing runs. Comparison of the non-redundant sequences with correct invariant ends and insert lengths from each sequencing run (sequence sets P1 and P2, with 36,031 and 64,600 such sequences, respectively) revealed that 7221 sequences were present in both samples. Using the equation described above, we calculated that the TJC-5 plasmid library has the potential to encode 322,338 different full-length sequences, approximately one-quarter the number of bacterial colonies pooled when it was generated. (If there were indeed 1.2 million different sequences in the TJC-5 library, the capture–recapture equation predicts that the overlap between P1 and P2 would be fewer than 2000 sequences.) Therefore, we conclude that many of the transformants pooled to generate the library contained redundant sequences or sequences that could not encode the intended protein product. Thus, simply counting the number of transformants substantially overestimates the number of different, full-length proteins encoded by the library.

In order to validate this approach for our particular application, we tested the capture–recapture equation in a series of *in silico* simulations in which we randomly sampled a library with a defined number of sequences, as described in the ‘Supplementary material’ section. This analysis suggested that this calculation is suitable for estimating the size of randomized libraries based on the number sequences in common between two independent sets of non-redundant sequences. Notably, sampling a small fraction of the library was sufficient to achieve an accurate estimate of its size (Supplementary Table S2).

**Composition and complexity of the library after infection and recovery from mammalian cells**

The TJC-5 plasmid library was designed to be packaged into retrovirus particles and expressed in mammalian cells. The effect of retroviral transduction on the complexity and composition of randomized expression libraries has not been thoroughly investigated (Kitamura et al., 2003; Hale et al., 2004; Neylon, 2004; Kuzmicheva et al., 2009). In particular, we were concerned that there may be a genetic bottleneck during retroviral packaging, infection or recovery that may significantly reduce the complexity of the library or alter its composition.

To determine whether the introduction of the library into mammalian cells and subsequent recovery perturbed its composition, size or diversity, TJC-5 plasmid DNA was packaged into retroviral particles in 293T cells, and the resulting virus stock was used to infect mouse BaF3 cells at a low multiplicity of infection. After selection for hygromycin resistance, PCR was used to recover integrated library sequences from genomic DNA, and two independent amplification products (R1 and R2) were subjected to 454 pyrosequencing. A total of 181,565 sequences were obtained, of which nearly 93% had the correct invariant ends. Of the
96,257 non-redundant sequences, over 75% had a 66-bp randomized segment (data not shown).

The nucleotide sequences of the non-redundant, full-length library members revealed that the recovered library retained the composition of the plasmid DNA library. The randomized segments were virtually identical in the two libraries in terms of nucleotide and amino acid ratios (Fig. 5, Table I, and Supplementary Fig. S1) and exhibited similar deviations from the expected nucleotide and amino acid composition at codons 12 and 13. In addition, less than 4% of sequences contained mutations at the third position of codons 12–30, and the fixed positions retained the wild-type sequence 99.7% of the time. Therefore, despite the mutagenic nature of retroviral reverse transcription and PCR-based recovery, relatively few mutations were introduced during virus production, infection and recovery. Finally, pairwise comparisons of the sequences from the recovered library revealed that they displayed the predicted level of diversity, with the peak of this distribution centered at a distance score of 28 (Fig. 7).

The two independent samples from the recovered library, R1 and R2, yielded 41,155 and 37,293 non-redundant sequences, respectively. The capture–recapture equation was used to estimate the complexity of the library after its introduction into mammalian cells and recovery. According to this calculation, the recovered library was estimated to express 315,477 different, full-length proteins, within 2% of the calculated size of the library prior to packaging and infection. These results suggested that there were no major genetic bottlenecks in the methods used to produce and recover the TJC-5 library from mammalian cells.

Although the results presented above demonstrated that the plasmid library and the recovered library were very similar in terms of their composition, level of diversity and overall size, it remained possible that they encoded different proteins. To determine whether most of the sequences present in the plasmid library were retained in the recovered library, we conducted pairwise comparisons of the non-redundant, full-length sequences in the four sets of sequences (P1, P2, R1 and R2) and determined the number of sequences in common between each pair, as described in the ‘Supplementary material’ section. This analysis revealed that the number of sequences shared between any two samples was very similar regardless of the origin of the samples (Supplementary Table S3), implying that the vast majority of the proteins encoded by the plasmid library were indeed retained in the recovered library. Therefore, despite the additional steps required to package the TJC-5 library into retroviral particles, express it in mammalian cells and recover sequences by PCR, the library recovered from cells was strikingly similar to the plasmid library in terms of its composition, level of mutagenesis, diversity, size and sequence content.

Summary
The results reported here validate the methods we used to construct, characterize and maintain complex expression libraries for use in mammalian cells. Deep sequencing allowed us to determine the size, composition and diversity of a randomized protein expression library and to demonstrate the absence of any significant bottleneck during the passage of the library through mammalian cells. As the cost of deep sequencing continues to decline and as the accuracy of these platforms further improves, this approach may prove to be an effective way to routinely determine the size and composition of similar libraries to optimize their use in protein engineering. Additionally, if adequate sequence coverage can be obtained, high-throughput sequencing can be used to identify the enrichment or depletion of specific library sequences during selection.

Supplementary data
Supplementary data are available at PEDS online.

Acknowledgements
We thank Bill Sugden and Hyonho Chun for material and helpful comments. DNA sequencing was conducted by the Yale Center for Genomics and Proteomics.

Funding
This work was supported by a pilot grant from the Yale Center for Genomics and Proteomics, a developmental grant from the Yale Skin Cancer SPORE (P50 CA 121974-02; R. Halaban, P.I.), and a grant from the National Cancer Institute (CA037157) to D.D. S.A.M. was supported by an institutional training grant (T32 GM007499).

References