Supporting Information for:

Deep sequencing analysis of phage libraries using Illumina

Wadim Matochko,[1] Kiki Chu,[2] Bingjie Jin,[1] Sam W. Lee,[2] George Whitesides[3], Ratmir Derda[1,*]

1. Department of Chemistry and Alberta Glycomics Centre, University of Alberta, Edmonton, Alberta, Canada, T6G 2G2.
2. Cutaneous Biology Research Center; Massachusetts General Hospital and Harvard Medical School; Charlestown, MA 02129 USA
3. Department of Chemistry, Harvard University, Cambridge, MA 02138, USA

Files not included here:

MatLab.zip An archive, which contains MatLab scripts, listed in Table S3

Raw.zip An archive, which contains smallFASTQ.txt. It is raw FASTQ file, representing 1/20th of the 6 Gb file. This file could be used to test the scripts.

Analysis.zip An archive, which contain the results of the analysis:
uniqueN_QF.txt Copy number analysis of peptides from forward reads
uniqueN_QR.txt Copy number analysis of peptides from reverse reads

Region from M13KE vector with variable insert

```
1   GTG GTA CCT TTC TAT TCT CAC TCT                          24
25  NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK NNK   60
61  GGT GGA GGT TCG GCC GAA                          78

Abbreviated as— TAT TCT CAC TCTR36GGT GGA GGT TCG
```

Primers

```
L1:    5'- NKK NKK ACT ATC TAT TCT CAC TCT -3'

R1:    5'- CGA ACC TCC ACC -3'

R2:    5'- TTC GGC CGA ACC TCC ACC -3'    (longer complimentary region)
```

(ACTATC is the 6-mer barcode)
(NKKNKK is a random hexamer that will facilitate cluster formation)

********** **primer alignment for L1+R1 pair** **********

```
5' NKK NKK ACT ATC TAT TCT CAC TCT                    3'   (L1)

5′                 TAT TCT CAC TCTR36GGT GGA GGT TCG 3′
3′                 ATA AGA GTG AGAR36CCA CCT CCA AGC 5′

3'                                 CCA CCT CCA AGC 5' (R1)
```

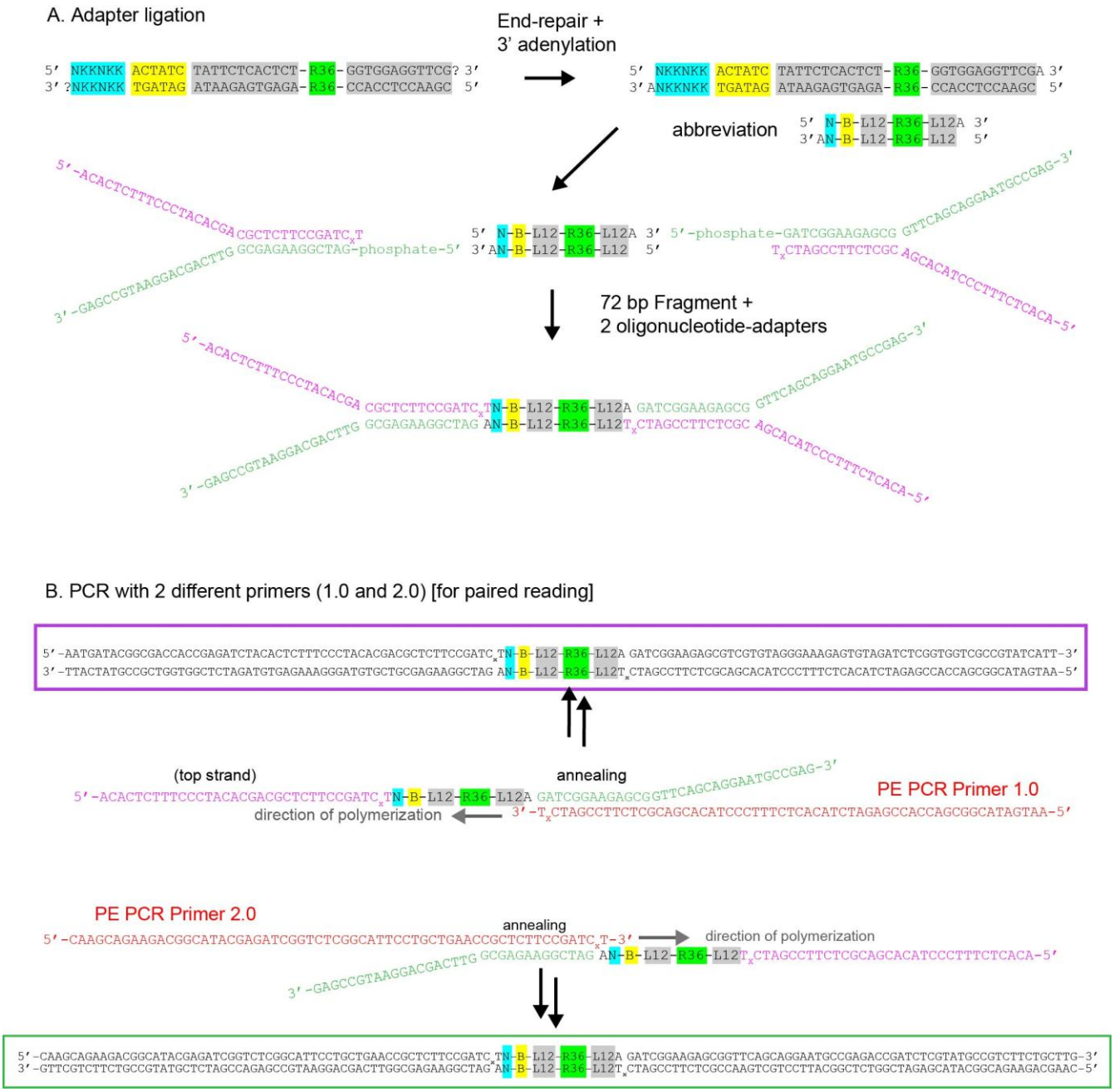
Result after PCR (72 bp fragment):

```
5′ NKK NKK ACT ATC TAT TCT CAC TCTR36GGT GGA GGT TCG 3′
3′ NKK NKK TGA TAG ATA AGA GTG AGAR36CCA CCT CCA AGC 5′
```

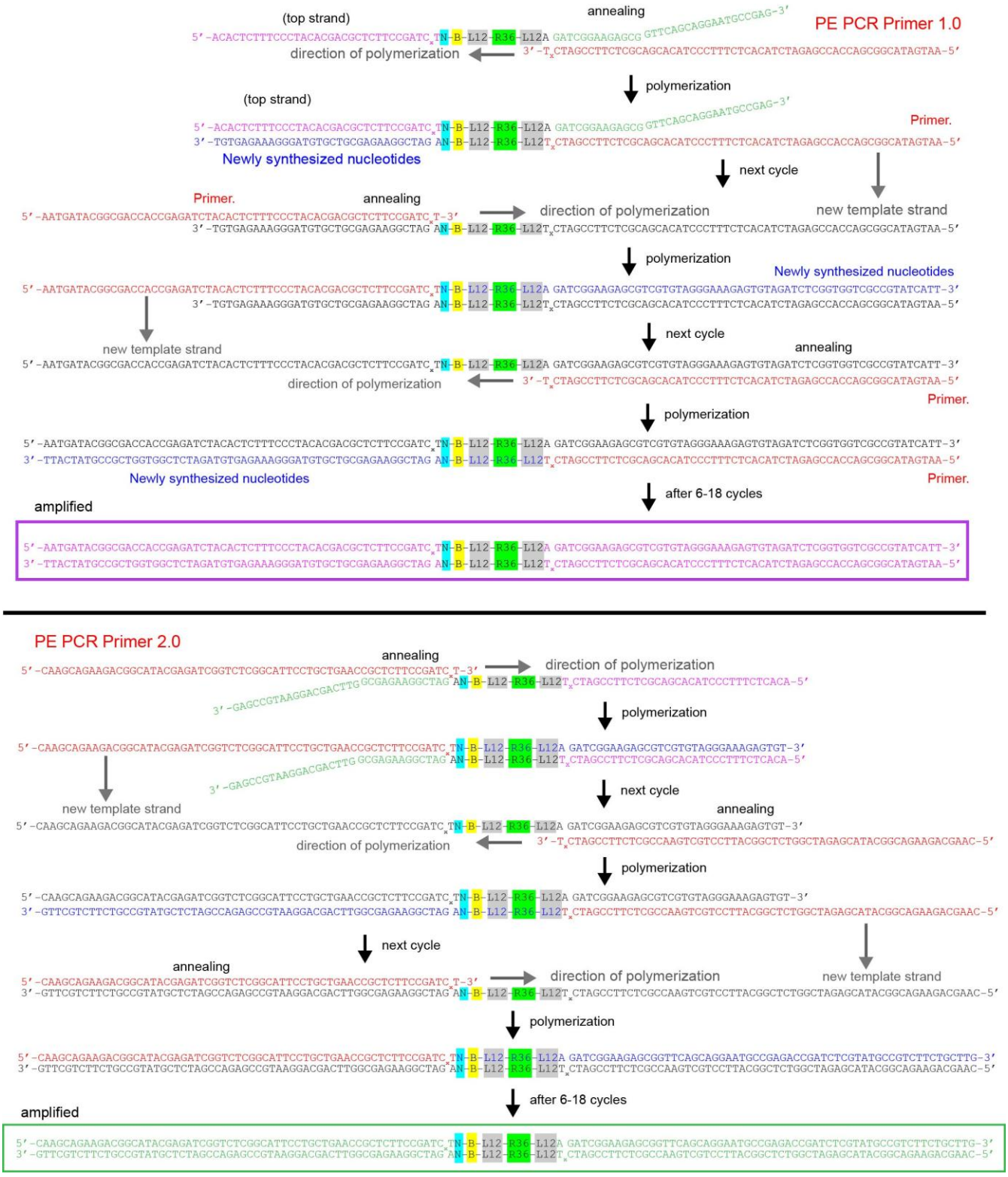
After ligation of the adapters and PCR (190 bp fragment):

```
5′ ILL-LEFT-NKK NKK ACT ATC TAT TCT CAC TCTR36GGT GGA GGT TCG-ILL-RIGHT 3′
3′ ILL-LEFT-NKK NKK TGA TAG ATA AGA GTG AGAR36CCA CCT CCA AGC-ILL-RIGHT 5′

ILL-LEFT- = 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCxT-
ILL-LEFT- = 3′-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGxA-

-ILL-RIGHT = -AxGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT-3′
-ILL-RIGHT = -TxCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA-5′
```

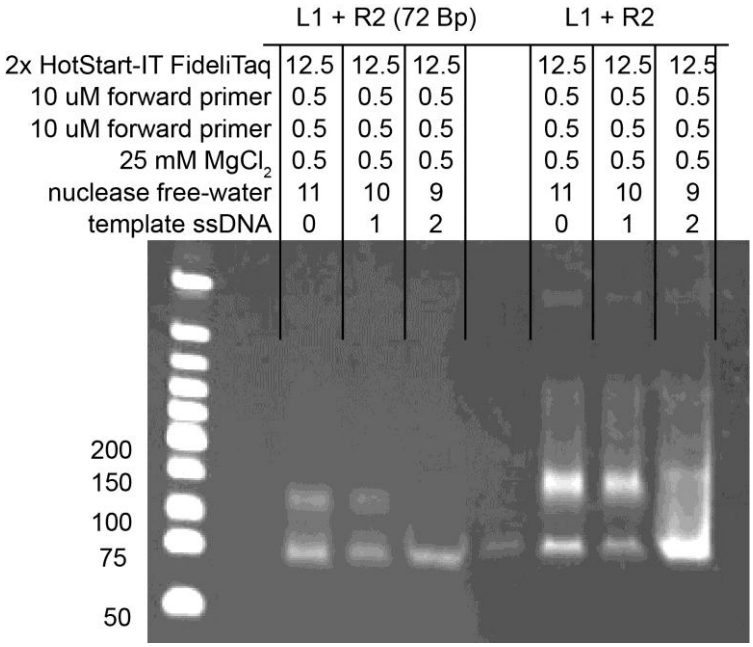
Scheme S1 First generation of the design of the primers. Alignment of primers to the (+) and (-) strands of the M13KE vector and expected products after PCR. We designed and tested two right primers: shorter R1 and longer R2. Both primers yielded expected product after amplification (see Supporting Figure 1). We selected the R1, because we were aiming to find the primer of the shortest possible length. For sequences of the adapters and PCR primers see Schemes S2, S3. Complete sequence of M13KE is available from New England Biolabs:
http://www.neb.com/nebecomm/tech_reference/restriction_enzymes/sequences/GenBank/M13KE.gbk.txt



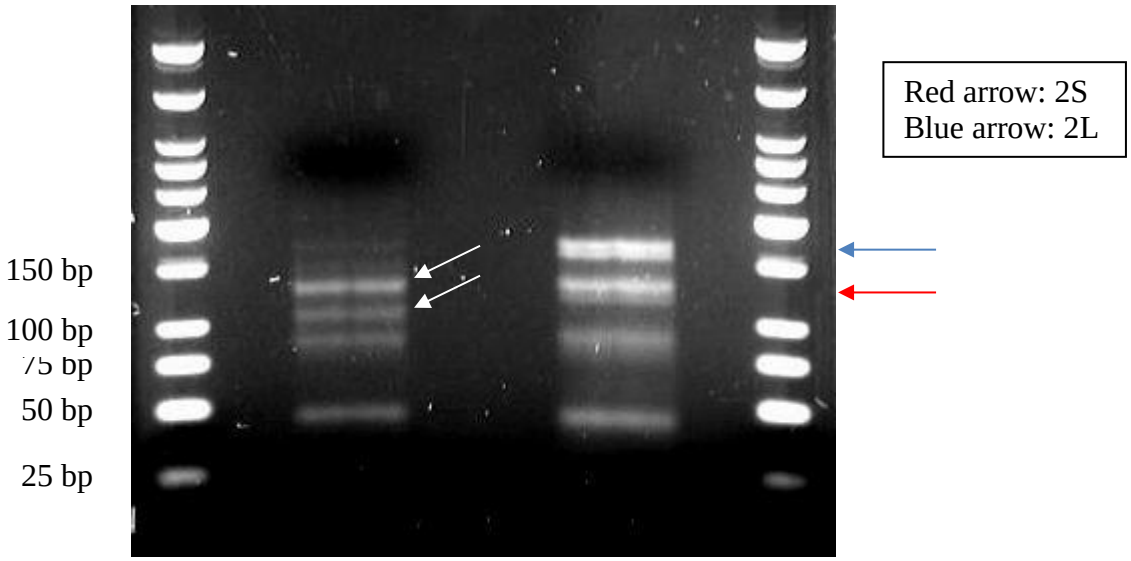
Scheme S2. Step-by-step description of the end-repair, 3'-adenylation, and ligation of paired-end (PE) Illumina adapters. Question marks at the 3'-end in the initial PCR product indicate a mixture of adenylated and non-adenylated ends. We observed significantly higher ligation efficiency after end-repair and re-adenylation or all 3'-ends. (B) Double stranded DNA after amplification with Illumina PE primers. For detailed steps of the PCR amplification see Scheme S3.



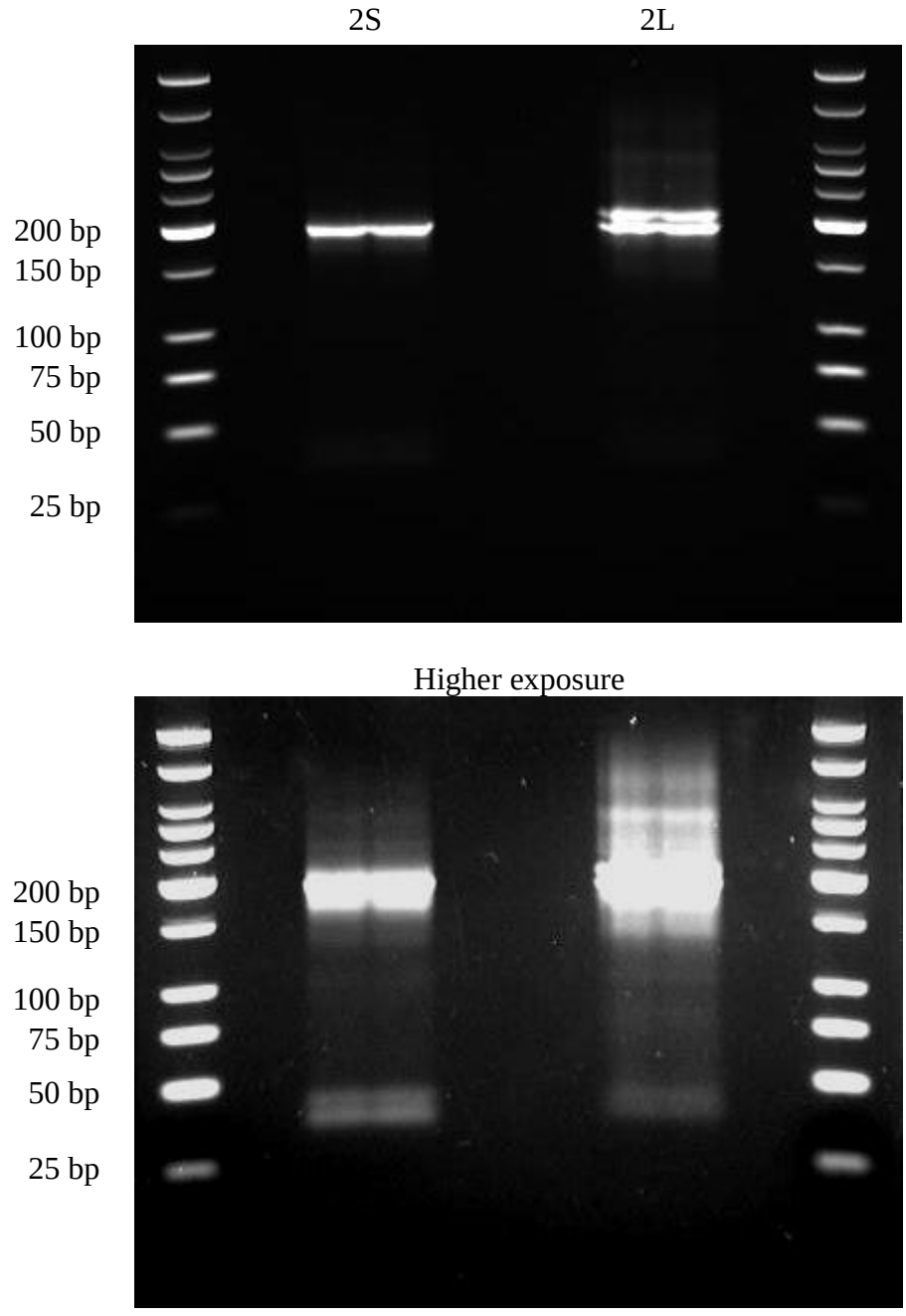
Scheme S3. PCR-amplification of the ligated products. C_xT indicates phosphorothioate bond between C and T. Sequences of the adapters and primers were supplied by Illumina.



Supporting Figure S1. PCR of the M13KE template with L1/R2 or L1/R2 primer pairs. At low concentration of template, primers yield non-specific product (potentially a primer dimer). Products were visualized on precast 10% TBE acrylamide gel (Invitrogen, Cat# EC62752BOX)



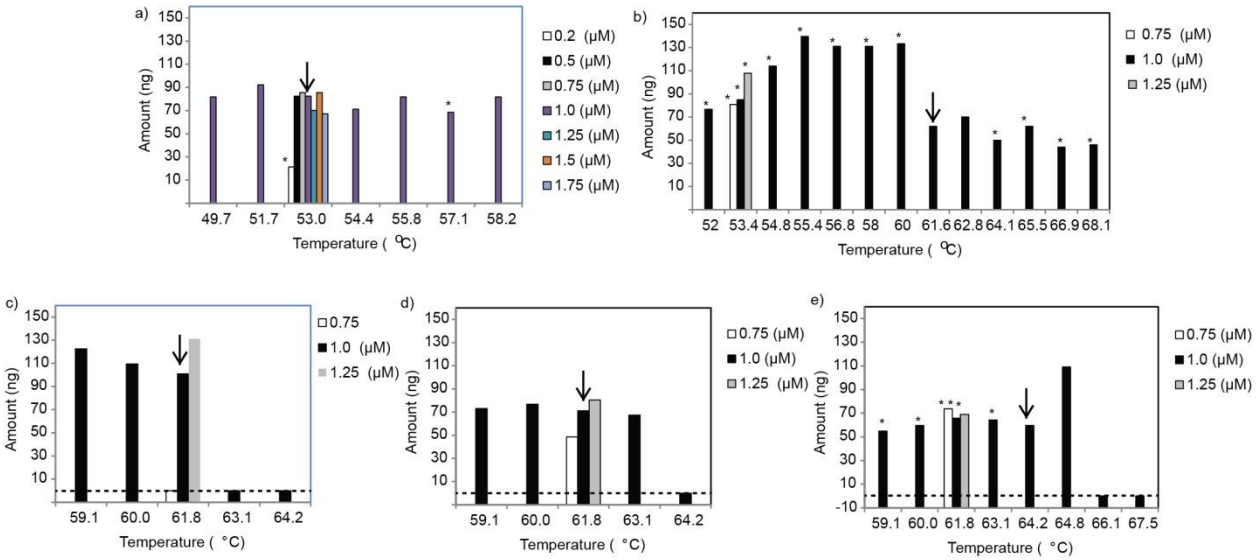
Supporting Figure S2. Adapters (33bp x 2) were ligated to the 72bp fragment and purified on the gel. Left: No end repair with 18 hour ligation at 16°C. Right: With end repair → 3' adenylation → 30 min ligation at 20°C. In the sample 2, the red arrow indicated the expected fragment (~140bp). The blue arrow fragment could also be the desired ligation product. In the Illumina protocol, it is often referred to as "back up" library. These two bands were excised purified separately and enriched by PCR (Supporting Figure 3).



Supporting figure S3. 2% agarose gel was ran after PCR amplification with primers that compliment the adapters in order to enrich the DNA fragment that is successfully ligated with the adapters. The final product should be about 200 bp—the PCR adds about another 30 bp on each side. We observed two products (188bp and 192 bp). The fact that both 2S and 2L amplified suggests that both fragments contain the adapters. Both 2S and 2L were excised and purified. The DNA was concentration in the purified sample was validated by Qbit.

```
5'- NKKNBARTAT TCT CAC TCT -3'   (left-BAR1-NKKN)

5′- NKKNBARCGA ACC TCC ACC -3′   (right-BAR1-NKKN)

********** primer alignment ******************

5'- NKKNBARTAT TCT CAC TCT -3'   (left)

5′          TAT TCT CAC TCT(NNK)12GGT GGA GGT TCG -3′

3′          ATA AGA GTG AGA(NNK)12CCA CCT CCA AGC -5′

3'                              CCA CCT CCA AGCRABNKKN -5'(right)

**************** fragment ********************

5′- NKKNBARTAT TCT CAC TCT(NNK)12GGT GGA GGT TCGRABNKKN -3′

3′- NKKNBARATA AGA GTG AGA(NNK)12CCA CCT CCA AGCRABNKKN -5′
```

Scheme S4. Second generation of the design of the primers for the amplification of phage libraries. Each primer contains the NKKN region, barcode region (BAR), randomized nucleotide sequence ($(NNK)_{12}$) which corresponds to random amino acid sequence for Ph.D-12™ phage, compliment sequence at the 5' end of the variable region (TAT TCT CAC TCT), and the reverse compliment at the 5' end (CCA CCT CCA AGC). Different primers used were only different at the BAR code region only.



Supporting figure S4. Optimization of PCR conditions for BAR1-BAR5 primers. The optimization of BAR 1 is represented in plot (**a**), BAR 2 in plot (**b**), BAR 3 in plot (**c**), BAR 4 in plot (**d**), and BAR 5 in plot (**e**).

* Illustrates the concentration and temperature that resulted in multiple bands.

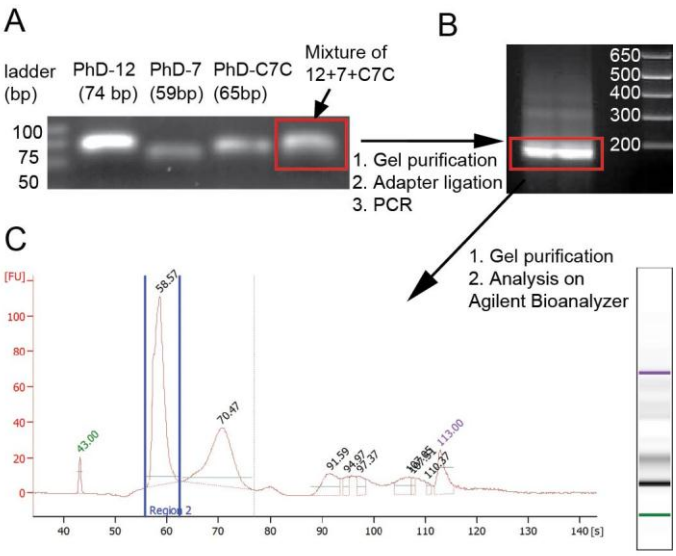
→ Indicates the condition chosen for PCR. We selected conditions, which yielded consistent amount of dsDNA blunt ended product and yielded no multiple bands on the 2% agarose gel.

	BAR 1	BAR 2	BAR 3	BAR 4	BAR 5
H_2O	24.5 µL				
5x Phusion Buffer	10 µL				
10 mM $MgCl_2$	2.5 µL				
10mMdNTPs	1 µL				
10 µM left-BAR	5 µL				
10 µM right-BAR	5 µL				
M13 phage DNA template (50 ng/µL)	1.5 µL				
Phusion Hot Start DNA polymerase (2U/µL)	0.5 µL				
During amplification.					
Step 1 → 30sec	98°C				
Step 2 → 10sec	98°C				
Step 3 → 20sec	53°C	62°C	62°C	62°C	64°C
Step 4 → 30sec	72°C				
Step 5 → Repeat Steps 2-4, 34x					
Finish → 72°C 5min					
Hold → 4°C					

Table S1. The optimized conditions for each primer used in the amplification of each variable region of the phage library.

Primer	Right Primer		Left Primer		T_{anneal} (ºC)
		T_m (ºC)		T_m (ºC)	
BAR1	NKKN GTA CGA ACC TCC ACC	55.9	NKKN GTA TAT TCT CAC TCT	46.1	53.0
BAR2	NKKN GAC CGA ACC TCC ACC	58.7	NKKN GAC TAT TCT CAC TCT	48.7	62.0
BAR3	NKKN TTG CGA ACC TCC ACC	57.2	NKKN TTG TAT TCT CAC TCT	47.0	62.0
BAR4	NKKN TCA CGA ACC TCC ACC	53.4	NKKN TCA TAT TCT CAC TCT	46.5	62.0
BAR5	NKKN CGA CGA ACC TCC ACC	59.0	NKKN CGA TAT TCT CAC TCT	49.2	64.0

Table S2: Sequences and melting temperatures of each primer. T_{anneal} is the optimal annealing temperature selected for each set of primers.



From [s]	To [s]	Corr. Area	% of Total	Average Size [bp]	Size distr. in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]
55.77	62.50	438.4	46	196	6.8	1,401.89	10,873.9
62.66	76.91	313.5	33	327	10.8	861.46	4,054.0

Supporting figure S5. Processing multiple libraries and multiple experiments using barcoded primers. (A) PCR Isolation of variable regions from three different libraries: 12-mer library (Ph.D.-12, New England Biolabs); 7-mer library (Ph.D.-7, New England Biolabs), cyclic 7-mer library (Ph.D.-C7C, New England Biolabs). Each library gives a single band on a gel. After PCR isolation, the libraries can be mixed together and processed as a mixture. (B) 2% agarose gel describing results after gel purification, end repair, adapter ligation and PCR amplification. Strong band at ~200 bp was excised from the gel. The product was analyzed on Agilent Bioanalyzer prior to Illumina sequencing. (C) Agilent trace and "gel view" or the trace (on the right). The table on the bottom describes molecular weight and concentrations for two major peaks.

#	Script name	Brief description of function
1	runALLscripts.m	Executes scripts #2, #3, #4 and #5 for one or several FASTQ files
2	rawseq.m	Breaks FASTQ file into 250,000-line text files The files rawseq0001.txt to rawseq0129.txt are saved in a directory XXX_RAW, where XXX is the name of the FASTQ file.
3	parseq.m	Opens raw files created by script #3, searches for forward and reverse adapter sequences and breaks raw text file into functional sequence regions. The files: parseqF0001.txt - parseqF0129.txt parseqR0001.txt - parseqR0129.txt parseqREM0001.txt - parseqREM0129.txt Are saved in a directory XXX_PAR, where XXX is the name of the FASTQ file. "F" and "R" designate files in which the sequences were identified using forward or reverse adapters. "REM" files contain sequences that could not be mapped
4	quaseq.m	Opens mapped sequence files created by script #4, analyzes the quality and NNK-format of each sequence and translates sequences to amino acids. The files: quaseqQF0001.txt - quaseqQF0129.txt quaseqQR0001.txt - quaseqQR0129.txt Are saved in a directory XXX_QUA, where XXX is the name of the FASTQ file. "F" and "R" designate files in which the sequences were identified using forward or reverse adapters. The script also could also save erroneous reads, which could be repaired: quaseqEF0001.txt - quaseqEF0129.txt quaseqER0001.txt - quaseqER0129.txt These files are small (<5% of QF and QR files) and contain sequences similar to those found in QF and QR files.
5	uniseq.m	Opens sequence files created by script #4, and analyzes the frequency of unique nucleotides. Only NNK-sequences are processed. The results are saved into four files in a directory XXX_UNI, where XXX is the name of the FASTQ file High-quality reads: uniqueQF.txt and uniqueQR.txt Repaired erroneous reads: uniqueEF.txt and uniqueER.txt
6	truncseq.m	(Optional): Processes files created by script #5 and truncates 12-mer sequences into 11-mer, 10-mer, etc, and re-analyzes the frequency of truncated files. The script writes eleven files, for each file created by script #5, and saves them in directory XXX_UNI under names: uniqueQF(12toN).txt, uniqueQR(12toN).txt, etc The number N ranges from 1 to 11.
7	uniquef.m	Satellite frequency-analysis script used by scripts #5 and #6.
8	savePar.m	Satellite formatting script used by script #3
9	isNKK.m	Satellite script used by script #4. Analyzes the format of the sequence
10	makename.m	Satellite formatting script used by scripts #2-#6
11	uniqueCOMB.m	Satellite frequency-analysis script used by scripts #5 and #6.

Table S3: List of MatLab scripts used for processing of FASTQ files

(A) **Script: rawseq.m** 2h 28 min	**Sample time log for the processing:** Mac Air (laptop), 2.13 GHz Intel Core 2 Duo, 4 GB 1067 MHz DDR3 RAM, Mac Os X 10.6.8 file: s_1_sequence.txt 6.07 GB wrote rawseq0001 Load=46s Convert=0.45s Save=13s wrote rawseq0002 Load=56s Convert=0.71s Save=13s … wrote rawseq0129 Load=55s Convert=0.77s Save=12s
(B) Script: parseq.m 2h 43 min	rawseq0001.txt F:17.0s R:31.1s N:4.69s Mut:13.0s Trun:4.55s Rem:0.14s rawseq0002.txt F:17.1s R:31.3s N:4.24s Mut:11.9s Trun:4.33s Rem:0.166s … rawseq0129.txt F:17.8s R:30.5s N:4.48s Mut:13.0s Trun:4.83s Rem:0.191s
(C) Script: quaseq.m 1h 23 min	parseqF0001.txt 55484/85820 seq (64.7%). Rescued 0 seq (0%). Time: 11.8 sec parseqF0002.txt 58061/86134 seq (67.4%). Rescued 0 seq (0%). Time: 12.3 sec … parseqF0129.txt 68524/90520 seq (75.7%). Rescued 0 seq (0%). Time: 14.6 sec parseqR0001.txt 121852/162056 seq (75.2%). Rescued 0 seq (0%). Time: 25 sec parseqR0002.txt 136798/161763 seq (84.6%). Rescued 0 seq (0%). Time: 27 sec … parseqR0129.txt 139021/157190 seq (88.4%). Rescued 0 seq (0%). Time: 25.4 sec
(D) Script: uniseq.m 0h 47 min	quaseqQF0001.txt loaded. Found 22476 nuc in 52794 reads in 1.35 sec quaseqQF0002.txt loaded. Found 37023 nuc in 107998 reads in 1.57 sec quaseqQF0003.txt loaded. Found 48962 nuc in 164152 reads in 1.61 sec quaseqQF0004.txt loaded. Found 59797 nuc in 224116 reads in 1.8 sec … quaseqQF0129.txt loaded. Found 652813 nuc in 7902401 reads in 4.59 sec Wrote 652813 unique seq in 644 sec **Total time: 7h 21 min**

Scheme S3 Processing of a typical FASTQ file by MatLab scripts. The scripts generate detailed output, which outlines the processing time for each step. Although we optimized the script to minimize the processing time, we believe some steps could be further optimized.

(A) The script *rawseq* breaks the 6 GB FASTQ file into 129 files (rawseq0001.txt to rawseq0129.txt). Loading, conversion and saving time for each file is noted.

(B) The *parseq* script loads files rawseq000N.txt files and searchers for forward (F), reverse (R) adapters, as well as adapters with unknown nucleotides (N), mutations (Mut), truncations (Trun) and saves unidentified, remaining sequences (Rem). Time (sec) used for each search is noted.

(C) The *quaseq* script loads parseqF000N.txt and parseqR000N.txt files and assesses the quality of sequences. The output indicates the number of high-quality sequences/total sequences. The script could also rescue low-quality reads with errors in N*3rd position, and saves those sequences that could be unambiguously translated. The rescued sequences comprise ~1.5% of total sequences, and are similar to high-quality reads. The rescue option was turned "off" for this particular run. If this option is turned one, the rescued sequences are saved in separate file with tags "EF" (erroneous forward) or "ER" (erroneous reverse).

(D) The *uniseq* script loads quaseqQF000N.txt files (high-quality forward reads) and identifies unique nucleotide sequences. The nucleotides are then translated to peptides and all sequences are saved in uniqueQF.txt file. The same steps are repeated for quaseqQR000N.txt, (high-quality reverse reads). For clarity, the output for these files is not shown.

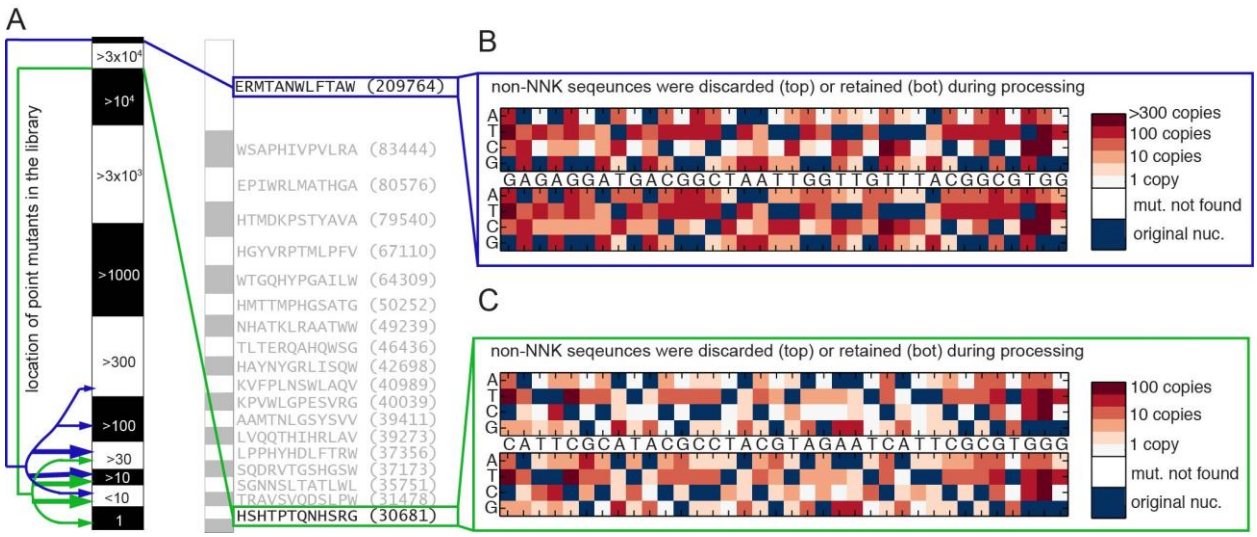


Figure S6. Analysis of point mutations in the library. We selected two abundant sequences (the most abundant and 19th most abundant), generated point mutations of these sequences and searched for these point mutations in the library. The approximate locations of these sequences in the library is showed by green and blue arrows. The size of the arrows qualitatively indicate abundances or mutants in each region.
(B) and (C) indicate positional abundance for each mutation. For example, the copy number of nucleotide that has G to C substitution in the 1st position is ~30; whereas abundance of G to T mutation in the same position is >300. We find significantly more point mutations than one would expect to have in sparse library (total number of clones is 10^6 while potential diversity is 10^{18}). In fact, for the top sequence we find all possible point mutations, including K to M mutations in positions 3, 6, 9 etc. To see these mutations, we re-analyzed the library and included non-NNK sequences in our analysis (see bottom heat plots in B and C)

The median abundance of point mutations is ~200 (B) and ~20 in (C), which is ~0.1% of the abundance of their original sequences. We concluded that most point mutations, thus, correspond to sequencing errors. We could thus assume that the region of the library that has abundances of >100 is free of sequencing errors.