



Deep sequencing analysis of phage libraries using Illumina platform

Wadim L. Matochko^a, Kiki Chu^b, Bingjie Jin^a, Sam W. Lee^b, George M. Whitesides^c, Ratmir Derda^{a,*}

^aDepartment of Chemistry and Alberta Glycomics Centre, University of Alberta, Edmonton, Alberta, Canada T6G 2G2

^bCutaneous Biology Research Center, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA 02129, USA

^cDepartment of Chemistry, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:

Available online 20 July 2012

Keywords:

Deep sequencing
Phage display
Peptides
Illumina
Amplification
Diversity

ABSTRACT

This paper presents an analysis of phage-displayed libraries of peptides using Illumina. We describe steps for the preparation of short DNA fragments for deep sequencing and MatLab software for the analysis of the results. Screening of peptide libraries displayed on the surface of bacteriophage (phage display) can be used to discover peptides that bind to any target. The key step in this discovery is the analysis of peptide sequences present in the library. This analysis is usually performed by Sanger sequencing, which is labor intensive and limited to examination of a few hundred phage clones. On the other hand, Illumina deep-sequencing technology can characterize over 10^7 reads in a single run. We applied Illumina sequencing to analyze phage libraries. Using PCR, we isolated the variable regions from M13KE phage vectors from a phage display library. The PCR primers contained (i) sequences flanking the variable region, (ii) barcodes, and (iii) variable 5'-terminal region. We used this approach to examine how diversity of peptides in phage display libraries changes as a result of amplification of libraries in bacteria. Using HiSeq single-end Illumina sequencing of these fragments, we acquired over 2×10^7 reads, 57 base pairs (bp) in length. Each read contained information about the barcode (6 bp), one complementary region (12 bp) and a variable region (36 bp). We applied this sequencing to a model library of 10^6 unique clones and observed that amplification enriches ~ 150 clones, which dominate $\sim 20\%$ of the library. Deep sequencing, for the first time, characterized the collapse of diversity in phage libraries. The results suggest that screens based on repeated amplification and small-scale sequencing identify a few binding clones and miss thousands of useful clones. The deep sequencing approach described here could identify under-represented clones in phage screens. It could also be instrumental in developing new screening strategies, which can preserve diversity of phage clones and identify ligands previously lost in phage display screens.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Phage display is a powerful method for the discovery of peptides that bind to any target [1,2]. The binding of phage library to a target, or “panning”, narrows the naïve library of 10^9 clones to 10^5 – 10^6 clones. This is a typical number of phage clones recovered after one round of panning, but only some of these clones have affinity for the target. To narrow the diversity of true binding clones the library is amplified in bacteria. Amplification multiplies the copy number of each clone and generates a focused library, which can be panned again [2]. Rounds of panning and amplification narrow the diversity of the library and enrich for phage clones that present target-binding peptides. The key to this process is the analysis of peptide sequences present in the library at various steps of the screening. Sequences enriched as a result of selection correspond to the specific binders against the target. Conventional San-

ger sequencing of clones require isolation of DNA from individual phage clones. It is a labor-intensive process and is rarely used to analyze more than a hundred library clones.

Analysis of a small number of sequences enriched in a screen can be used to predict one consensus motif [3,4]. Phage-display screens could also yield a large number of consensus motifs. For example, thousands of diverse sequence motifs should emerge from the panning against intact cells because an average cell contains thousands of structurally diverse receptors. If a screen selects a large number of independent binding clones, one has to sequence large numbers of clones to identify all the useful binding sequences. Arap, Pasqualini and co-workers were the first to use 454 sequencing to analyze $\sim 50,000$ sequences from a library of 7-mer peptides; the author applied this technology to identify peptides emerging from panning against different organs *in vivo* [5]. Subsequently, the same sequencing was used by several groups to monitor selection of binding proteins from a library of open reading frames (ORF) displayed on phage [6,7]. Sidhu and co-workers used 454-sequencing to boost selection of peptides binding to

* Corresponding author.

E-mail address: ratmir.derda@ualberta.ca (R. Derda).

different PDZ domains [8–10]. Notably, the authors used barcoded primers for the preparation of a library for sequencing and, thus, sequenced 22 independent panning experiments in one run. [8] Lerner and co-workers applied 454 sequencing to find antibodies that bind to various proteins displayed on the surface of bacteria [11]. Sequencing technologies of throughput higher than 10^4 – 10^5 could provide more complete coverage of the libraries. Increased throughput could also allow analysis of multiple experiments in a single run. Illumina/Solexa deep-sequencing technology analyzes a library of blunt-ended double stranded DNA (dsDNA) fragments and generates up to 10^9 base pair (bp) reads in a single run. For example, Fisher and co-workers recently demonstrated the use of Illumina sequencing to characterize phage-displayed libraries of single chain antibodies (scFv) [12]. Fields and co-workers used Illumina sequencing to characterize selection from libraries of WW-protein displayed on T7 phage [13]. Johan den Dunnen and co-workers used Illumina to characterize peptide libraries after one round of panning against cell surface receptors [14]. In this paper, we present a one-step PCR that converts a library of M13KE plasmids isolated from the phage library to a collection of short dsDNA sequences suitable for Illumina sequencing. Using custom MatLab software, we perform large-scale analysis of sequence diversities.

Using deep sequencing, we explore the effects of amplification of phage libraries in bacteria on the diversity of peptides in these libraries. In previous publications, the result from sequencing of ~ 100 phage clones suggested that the amplification process enriches for specific peptide sequences [15,16]. Large scale sequencing, however, can provide observations that could not be interpreted from the sequences of 100 clones [17]. For example, deep sequencing of a library of DNA aptamers demonstrated that repeated amplification does not select for particular sequences. Instead, it enriches DNA sequence motifs that have low stability [18,19]. In this report, we analyzed diversity of amplified libraries using Illumina and observed a collapse of diversity in phage-displayed libraries after a single round of growth in bacteria. The collapse of the 10^6 -scale library to a few hundred abundant sequences would not be visible in small-scale Sanger sequencing [17,20,21]; it could also have been difficult to detect with smaller-throughput 454 Sequencing.

Characterization of sequence diversity is important for phage display technology, which has been used in over 5000 publications and patents in the past 20 years. It has enabled the discovery of ligands for hundreds of targets, yet the literature still contains several poorly-explained observations: (1) identical sequences could emerge from unrelated screens for unrelated targets [22,23], and (2) screens that should yield a large number of diverse ligands often yield only one sequence motif (reviewed in [16]). The nearly complete sequence coverage of libraries illuminates the origin of these observations. It highlights that the collapse of diversity in amplification might be one of the major limitations of phage-display technology. Deep-sequencing analysis will make it possible to bypass problems originating from the unwanted collapse of diversity [11]. Large-scale analysis could also help develop methods that preserve diversity of peptide libraries [24,25]. It can be used to enable discoveries of ligands that previously have been lost in phage display screens.

2. Experimental design

2.1. Choice of the library

In this report, we sequence a commercially-available library of random 12mers from New England Biolabs (Ph.D-12). This library has been used in ~ 800 publications (source of estimate: PLoMics database <http://www.treeofmedicine.com/phagedisplay> and Mi-

moDB database [22,23]). According to the manufacturer (NEB), naïve library contains up to 10^9 different sequences. Since this number is beyond the sequencing capabilities of Illumina, we worked with 1/1000th portion of the library containing 10^6 different sequences. If a sequencing run produces 20 million sequences, the observed frequency of sequences could be approximated by a Poisson distribution with an expectation value of 20. For the above uniform library of 10^6 clones, the distribution predicts that every sequence will be observed at least 5 times. Over 99% of the library should be observed within 3 standard deviation of the expectation value ($\sqrt{20} \times 3 = 13$). The majority of the clones, thus, should be present at 7–33 copies.

To explore the effect of amplification on library diversity, we amplified a pool of 10^6 clones to 10^{13} pfu and isolated ssDNA from the combined pool of phage. Approximately 10^8 copies of each clone should be present in this pool. If relative abundances of clones were not changed during amplification, abundances of clones observed after deep sequencing should follow the Poisson distribution described above. In reality, we observed that a distribution of clones was dramatically different from the Poisson distribution, suggesting that growth preference of individual clones led to enrichment of some clones and depletion of others.

3. Description of materials

3.1. Isolation of DNA from phage libraries

3.1.1. Reagents

Polyethylene glycol MW 8,000 (PEG) (Fisher BP233-1), sodium chloride (NaCl) (Fisher S271-500), chloroform (Sigma 319988), phenol (Fisher A9311-1), anhydrous ethanol, distilled water, sodium iodide (NaI) (Fisher BP323-100), sodium acetate (Fisher S78229-1), glycogen (Invitrogen 10814-010).

3.1.2. Materials

1.7 Microcentrifuge tubes (Fisher 14222168), PEG/NaCl solution (20% (w/v) PEG/2.5 M NaCl, sterilized by autoclaving), micropipettes (Mandel P2N, P10N, P200N, P1000N) and micropipette tips (Fisher 02-707-439 (10 μ L), 02-707-430 (200 μ L), 02-707-404 (1000 μ)), benchtop microcentrifuge.

3.2. Preparation of the DNA for sequencing

3.2.1. Reagents

Hot start high fidelity DNA polymerase (e.g. Affymetrix Hot-Start-IT[®] Taq DNA Polymerase (71195) and Phusion[®] Hot Start II High-Fidelity DNA Polymerase (Finnzymes F-549L)), Illumina paired-end DNA sample prep kit (Illumina), QIAquick PCR purification kit (Qiagen 28104), QIAquick MinElute PCR purification kit (Qiagen 28004), QIAquick gel extraction kit (Qiagen (28704)), DNA loading buffer (50 mM Tris pH 8.0, 40 mM EDTA, 40% (w/v) sucrose), QIAEX II gel extraction kit (Qiagen 20021), certified low range ultra agarose (Bio-Rad 161-3106), $10\times$ TBE buffer (Bio Basic A0026), $50\times$ TAE buffer (Fisher FERB49), ethidium bromide (Fisher BP1302-10), DNA ladder (New England Biolabs N3233S), chloroform (Sigma 319988), phenol (Fisher A9311-1), anhydrous ethanol, distilled water, sodium acetate (Fisher S78229-1), glycogen (Invitrogen 10814-010).

3.2.2. Materials

1.7 ml microcentrifuge tubes (Fisher 14222168), DNA gel electrophoresis apparatus, micropipettes (Gilson P2N, P10N, P200N, P1000N), micropipette tips (Fisher 02-707-439 (10 μ L), 02-707-430 (200 μ L), 02-707-404 (1000 μ)), benchtop microcentrifuge, PCR thermal cycler.

4.3. Sequencing of the library

Concentration of dsDNA with ligated Illumina adapters was estimated using Qubit Fluorimeter (Invitrogen) or Agilent Bioanalyzer using manufacturer's protocol. The sample was diluted to the concentration of 10 nM and submitted for sequencing to Harvard FAS sequencing facility. The sequencing was performed using Illumina HiSeq and 50 bp single end reads.

4.4. Analysis of the library

- If FASTQ files are archived, extract plain text FASTQ files from archive
- Copy all MatLab files in one directory.
- Open “runALLscripts.m” file in MatLab editor and run it (F5).
- In the browse window, select one or several FASTQ files (hold “Shift”, to select several files at once).
- The program will test every file and display the first 10 lines from each file.
- If some files were selected incorrectly, or some files do not have FASTQ format, it is best to stop execution at this point (“ctrl” + “c”). The program will encounter error and stop when it encounters non-FASTQ file.
- If not interrupted, the program performs analysis of all selected files. A text-based output displays the progress of the analysis (see Supporting information Table S4 for example and explanation of the output).

5. Results

5.1. Isolation of variable dsDNA fragments from phage libraries

The majority of the phage display vectors share the same design: they contain a variable sequence flanked by constant regions containing restriction enzyme sequences (used for cloning of the library). We attempted to isolate the library sequences using KpnI and EagI restriction enzymes to isolate variable domains from M13KE vectors [26]. The collection of sticky-end fragments could be repaired to give blunt-ended fragments with identical termini. These fragments, however, could not be reliably sequenced by Illumina because the sequencing algorithm uses differences in terminal nucleotides to distinguish sequence clusters [27]. We attempted to introduce variable termini by ligation of short random nucleotide sequences; this approach, however, gave poor yields and was eventually abandoned. Nevertheless, we expect that excision by restriction nucleases could be useful for other deep sequencing approaches, such as Ion Torrent, which could process fragments with identical termini.

Our successful method for the isolation of variable regions used PCR amplification with primers complementary to the 12-bp constant regions flanking the variable sequence in the M13KE vector. The forward PCR primer contained a NKKNKK sequence at its 5'-position (Fig. 1A). Each primer, thus, was a mixture of $4 \times 2 \times 2 \times 4 \times 2 \times 2 = 256$ different primers. PCR with these primers generates dsDNA with 256 different blunt-end termini; this diversity should be sufficient for the algorithm that finds individual DNA clusters (colonies) during sequencing. We selected the NKKNKK sequence to minimize the possibility for hybridization with (NNK)₁₂ motifs in the library. The forward primer also contained a barcode sequence ATCACT. We selected this particular sequence after aligning all 256 (NKKNKK)-(ACTATC)-TATTCTACTCT sequences to (+) and (–) strand of M13KE vector. For all sequences, we observed hybridization of <7 bp, which should not interfere with PCR conditions optimized for 12 bp-long adapter sequences (Fig. 1C). We used a similar algorithm to find other barcode sequences (Supporting Information Tables S1 and S2). The use of

multiple barcodes allows for processing of multiple phage libraries in a single run (Supporting Information Fig. S5).

Successful PCR amplification of variable fragments was confirmed as a single band on 2% agarose gel. Amplification using primers with shorter variable regions or other barcode sequences yielded similar results (Supporting Table S1). Due to differences in melting temperatures of the primers, PCR conditions had to be re-optimized for each barcode sequence (Supporting Table S2). The fragments amplified from libraries of different size, such as 12-mer, 7-mer or 9-mer, gave dsDNA fragments of expected sizes. For example, the protocol described in Fig. 1A was validated using three different libraries: (1) Ph.D-12™, a library of 12-mers, 36 bp variable region; (2) Ph.D-7™, a library of 7-mers, 21-bp variable region, and (3) Ph.D-C7C™, a library of 7-mers flanked by Cys, 27 bp variable region. We used two primers with a total length of 38 bps long and observed PCR products close to the expected (1) 74, (2) 59, and (3) 65 bps (Supporting Fig. S5A).

5.2. Preparation of Illumina-compatible dsDNA Fragments

Ligation of DNA adapters that enable Illumina sequencing was performed according to the protocols supplied with the Illumina paired-end adapter Kit. Successful ligation of Illumina adapter sequences to the blunt-ended PCR product occurred only after end-repair of the product (Fig. 1D). Ligation yielded two products, referred to as **2L** and **2S**, with length similar to that of the expected product (140 bp for the 12-mer library). To enrich the DNA fragments, which were successfully ligated with the adapters, we run PCR amplification of purified **2L** and **2S** fragments with primers that complement the Illumina adapters. Both **2L** and **2S** yielded products of correct size after PCR (Fig. 1E) confirming that both **2L** and **2S** contained correctly ligated adapters. Both products were subjected to Illumina sequencing (single-read, 50 bp reads on HiSeq) yielding similar sequence abundances and diversities (see Fig. 3B below).

6. Overview of the analysis

6.1. Design of the analysis software

Sequencing by Illumina generates a ~4–10 Gigabyte text file. It is difficult to handle because, most desktop computers cannot open the file in a standard text editor. Additionally, Illumina is used primarily for genome sequencing, and most available software is written for assembly of genomes. Therefore, we wrote a software tailored for the analysis of phage libraries. The basic feature of the software is batch processing. The program first breaks the original 4–5 Gb FASTQ file into text files of ~100 Mb each. The subsequent processing, thus, requires less operational memory. Analysis proceeds in several steps: (i) conversion of one FASTQ file into smaller plain text files, (ii) identification of constant complementary regions and parsing of sequences, (iii) analysis of sequence quality, (iv) analysis of diversity of sequences, (v) translation of sequences, (vi) plotting. After each step, the program saves intermediate files in plain text (*.txt) format. Any intermediate text files can be opened and inspected in a standard text editor. Software written in MatLab was effective in analyzing a 4–5 Gb FASTQ file in 6–8 h on an average desktop or laptop computer (Supporting Information Scheme S5). We anticipate that re-writing the same script in a lower-level language (e.g. C++) could further accelerate the processing.

6.2. Overview of the scripts

Although the length of the dsDNA construct depicted in Fig 1C is 72 bp, single-end sequencing yielded reads of only 57 bp and con-

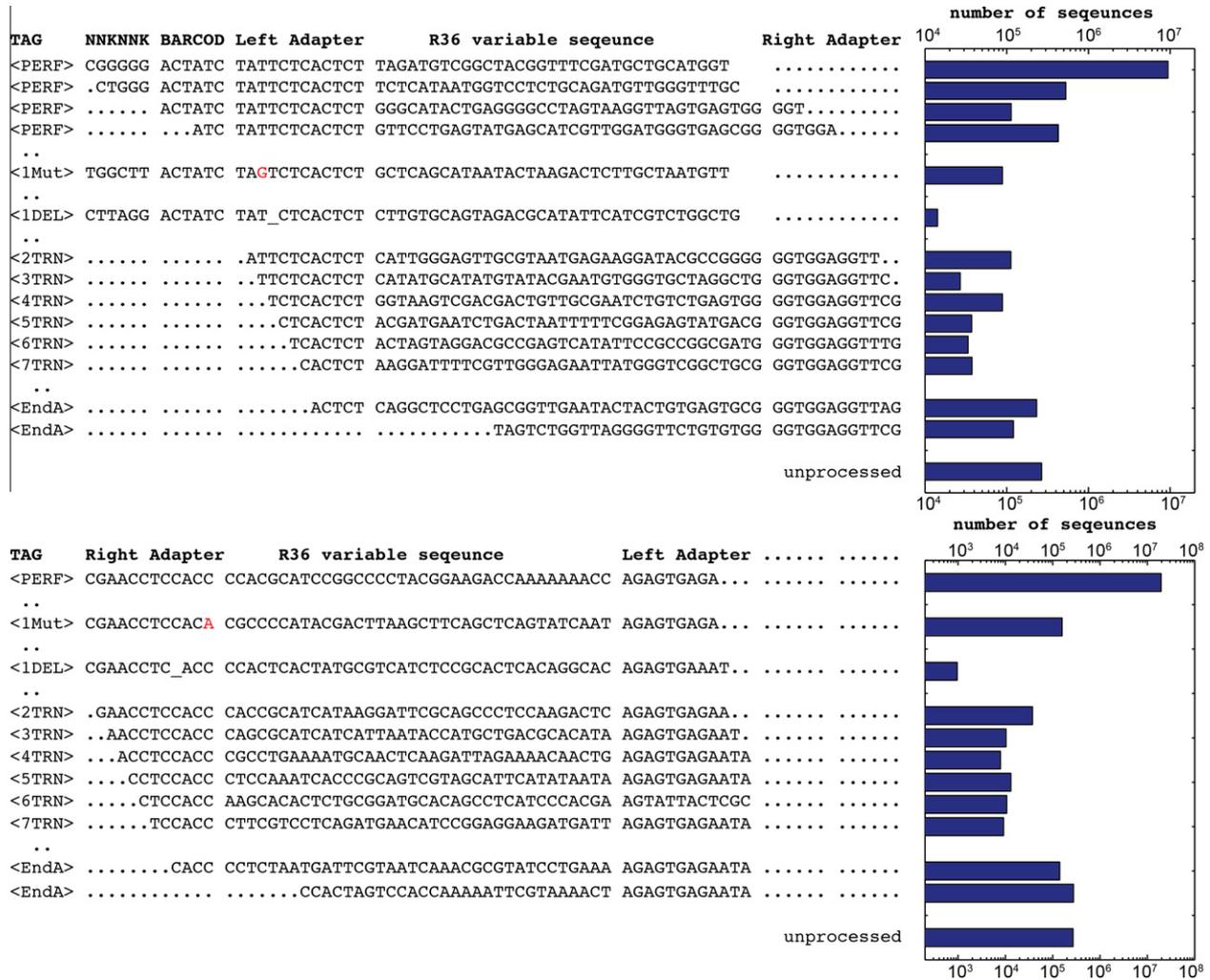


Fig. 2. Parsing of the full-length reads into mapped regions containing right adapter, left adapter, R36 variable region, NNK and Barcode regions preceding left adapter. Alignment was performed by searching for constant forward (A) or reverse (B) adapters. Tags at the beginning of each line describe the algorithm by which the adapter was identified. (PERF) perfect alignment; (1Mut) one mutation in the adapter; (1Del) one deletion in the adapter; (2TRN)(3TRN), etc. are truncation to 2nd, 3rd, etc. nucleotide in the adapter; (EndA) alignment to the adapter at the opposite end of R36 region. The log-scale plots on the right describe the relative abundance of sequences identified by specific algorithm.

tained complete sequence for only one constant region: either from the forward or the reverse primer. We designed the algorithm which used one constant adapter region to map the functional portions of the sequence: (1) NKKNNK portion, (2) barcode portion, (3) left adapter, (4) R36, and (5) right adapter (see Fig. 1A, C, F).

The process starts from the *rawseq.m* scripts, which breaks the original FASTQ file into smaller text files, 250,000 lines each. The *parseq.m* script then searched for the forward or the reverse adapter sequences (highlighted grey or blue in Fig 1C). We used a multi-step algorithm for the identification of the adapters. The majority of the sequences were mapped by perfect alignment to full-length adapter sequence ((PERF) in Fig. 2). 1% of sequences contained adapters with one mutation ((1Mut) in Fig. 2; mutation is highlighted in red). Few adapters had one internal deletion ((1Del) in Fig. 2; deletion is underscored, Fig. 2). A significant fraction of adapters had terminal truncations (lines tagged as (2TRN) to (7TRN) in Fig. 2). Truncated reads contained sequences of nucleotides from *i*th to (56 + *i*)th position (*i* = 2–25). Finally, primers with excessive truncations in one complementary region could be identified by alignment with the complementary sequence at the opposite end of the variable region (lines labeled as (EndA) in Fig. 2).

This algorithm mapped the majority of the forward and reverse reads (Fig. 2A forward and Fig. 2B for reverse search). Approximately 1.6% of sequences (0.5 million) could not be mapped because they contained a large number of low-quality reads or reads with multiple mutations or deletions in the adapter regions.

The parsed files were then processed by the *quaseq.m* script that assessed the quality of the R36 region containing the (NNK)₁₂ sequences. We selected only high-quality output in which all nucleotides had Phred Quality Score above 5 (this value could be changed in *quaseq.m* script on demand). High-quality sequences were then analyzed by *uniseq.m* script to generate abundances of nucleotides and cognate peptide sequences. The results were saved to *uniqueN_QF.txt* and *uniqueN_QR.txt* file (where F and R designate analysis of forward and reverse reads). The files are available as part of Supporting information.

In summary, from 32 million raw reads, the software identified ~11.1 million forward and 20.2 million reverse reads from which R36 sequences could be extracted. From R36 motifs with NNK structure, the software extracted 8.5 and 17.8 million peptide sequences from forward and reverse reads respectively. In current analysis of 12-mer libraries, the majority of the forward reads were

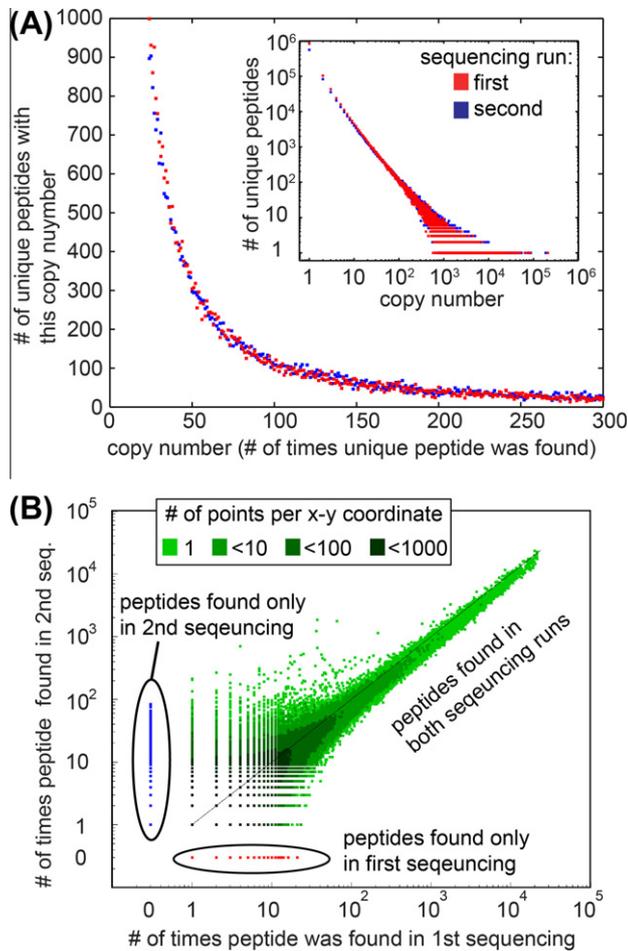


Fig. 3. (A) Abundance of peptides in the library; each point represents a peptide sequence. Red and blue colors represent two independent sequencing runs where red data correspond to **2S** and blue data correspond to **2L** library (Fig. 1D and E) prepared from the same amplified PhD-12 library. The insert describes a log–log plot of the same data. (B) Reproducibility of peptide abundances in two sequencing runs. The abundance of peptides at copy number >100 is highly reproducible between two runs. Peptides found in only run 1 (red dots) or run 2 (blue dots) have low relative abundance. Darker shades of green represent >10, >100 or >1000 data points in the same (x,y) coordinate.

truncated at the 11th amino acid (see *uniqueN_QF.txt*). Reverse reads, however, contained sequences for full-length 12-mer peptides (see *uniqueN_QR.txt*). We focused the remaining analysis on the 17.8 million reverse reads.

The script had options to retain or discard the sequences that did not have NNK format (i.e., sequences with A or C in position 3, or 6, or 9, etc.). If non-NNK sequences were retained, the results contained a significant fraction of sequences with TGA stop codons. M13KE vectors with stop codon in the N-terminal region of the *pIII* gene would lack N-terminal leader sequence and would not produce viable phage.[26] We concluded that TGA codons and other non-NNK codons are sequencing errors.

6.3. Preliminary analysis of sequence diversity in the library

Complete analysis of sequence diversities obtained using Illumina sequencing is beyond the scope of this manuscript. Here, we present the preliminary analysis of the sequences, and we confirm that sequencing runs are reproducible. Fig. 3 describes the distribution of sequence abundances in the library obtained by sequencing of two library preparations (bands **2S** and **2L** in Fig. 1D). The abundance of sequences in the two runs were similar

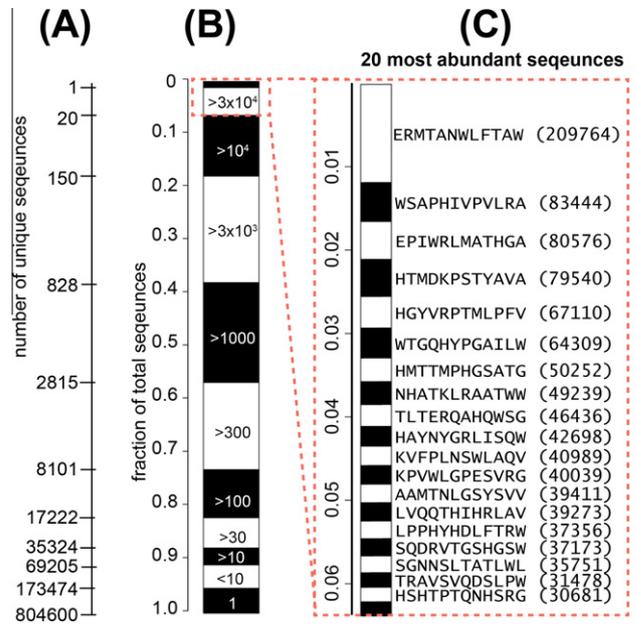


Fig. 4. Distribution of (A) the number of unique peptide sequences and (B) fraction of total peptide sequences in the library. Black-and-white stacked bar or “zebra-bar” describes the library in this and two subsequent figures (Figs. 5 and 6, S6). The height of each segment is proportional to the fraction that each sub-population occupies in the library. For example, ~5% of the library is occupied by 20 sequences, present at abundance of >30,000 copies. 20% of the library is occupied by 150 sequences, present at >10,000 copies, etc. (C) Zoomed-in zebra-bar describes top 20 sequences. The height of each segment is proportional to the fraction of the library occupied by each sequence. For example, top sequence occupies 1.2% of the library.

(see Fig. 3): some unique peptides were found in copy numbers of 10^4 and higher; nearly 10^6 peptide sequences were found in low copy number. The abundances of specific peptide sequences were highly reproducible between two runs (Fig. 3B). Peptides, which were observed 10^2 – 10^5 times in sequencing run 1, were observed at similar copy number in the 2nd sequencing run. Deviation from 1:1 correlation were observed at copy number <100. Some peptides, observed at copy number of 10–100 in the 1st run, were present at much lower copy number in run 2 or completely absent from the other sequencing run.

Distribution of sequence abundance was dramatically different from the predicted Poisson distribution with an expectation value of 20. It could not be modeled as Poisson distribution with any expectation value. A mere 20 clones constitutes 8% of the size of the library and were present at a copy numbers of >30,000 (Fig. 4). On the other hand, 500–800 thousand diverse sequences constituted another 8% and were present at a copy number of <10.

The distribution of sequence abundances followed the power-law distribution, producing a linear plot on a log–log scale (Fig. 3A, insert). We observed a deviation from this distribution for the low copy number peptides. Extrapolation of a log–log plot predicts that the number of single copy-number sequences should be 3 – 5×10^5 . The observed deviation suggested that a significant fraction of low-copy-number peptides could be the result of sequencing errors. Errors are abundant in Illumina sequencing [28], but we anticipate that many of these errors could be easily identified. One possible algorithm could be based on the assumption that the library is sparse. In other words, a library of nucleotides with structure $(NNK)_{12}$ has $(4 \times 4 \times 2)^{12} = 10^{18}$ members, and in a pool of 10^6 sequences, the probability to find a mutant is small. Despite this prediction, the search for point mutations of most abundant sequences yielded ~100 point-mutants for high-copy-number sequences (Supporting Information Fig. S6). The majority of these mutated sequences were present at low abundance (Figure S6); average abundance was ~1%, which is sim-

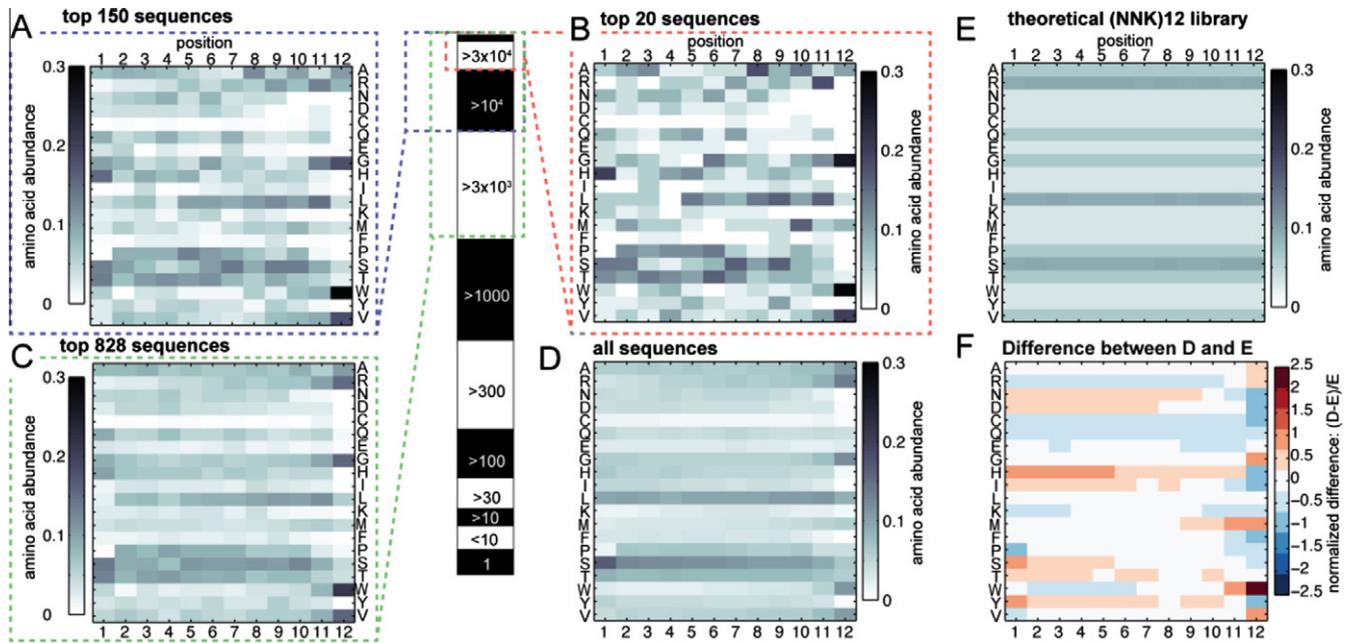


Fig. 5. (A–D) Positional abundance of amino acids in the top 20 sequences (B) is very different from abundance of amino acids in all peptides in the library (D). Abundance in top 150 sequences (A) and top 20 sequences (B) were similar. On the other hand, the abundance in the top 850 sequences (C) resembled that of the whole library. The sequences present at copy number of >30,000 are different from the rest of the sequences in the library. (D–F) Comparison of the distribution of the amino-acid in the entire library (D) and theoretical distribution of amino-acids in (NNK)12 library (E) reveals differences in positional abundances of individual amino acids. The plot in (F) describes fold-increase (red) or decrease (blue) in abundance of specific amino acids in specific position. Amino acid 1 corresponds to N-terminus and amino acid 12 to C-terminus.

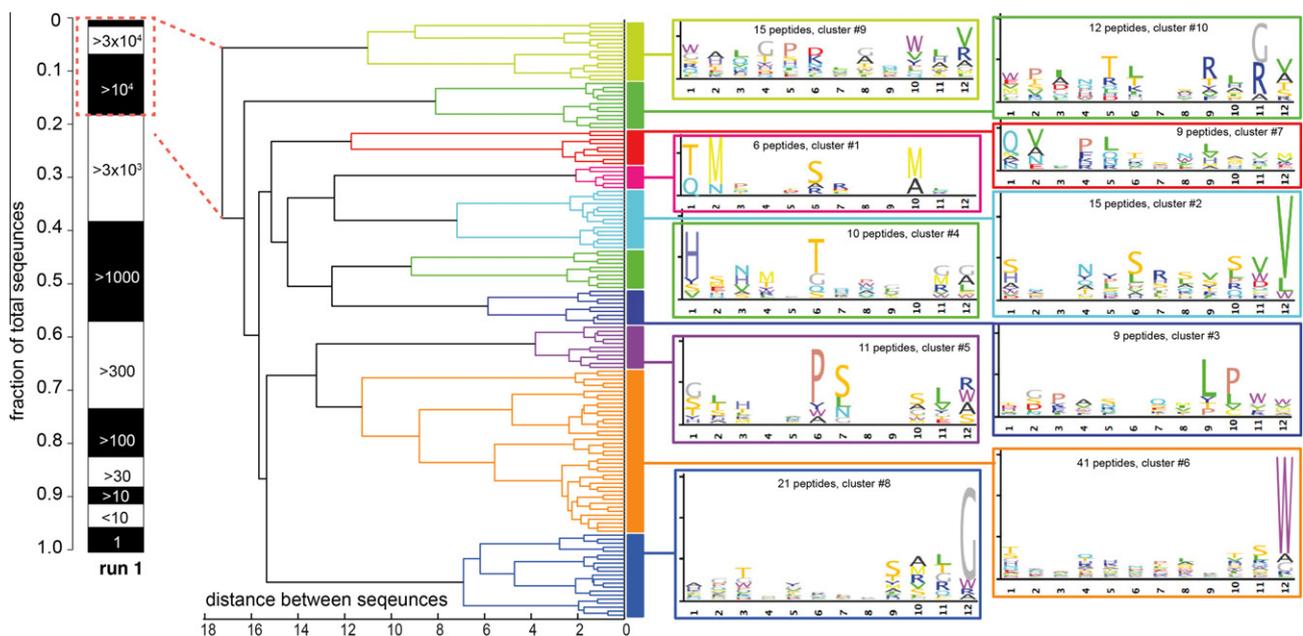


Fig. 6. Clustering analysis of the top 150 sequences (highlighted as dotted rectangle) based on sequence similarity. We observed 10 distinct clusters, which contained distinct consensus sequences. Calculation of distance and clustering was performed using Euclidian metric in MatLab (see Supporting information for scripts). Consensus motifs were generated using protein LOGO (pLOGO)[30]. Amino acid 1 corresponds to N-terminus and amino acid 12 to C-terminus.

ilar to the frequency of point mutations in adapter sequences (compare $\langle \text{PERF} \rangle$ and $\langle \text{1Mut} \rangle$ in Fig. 2). This preliminary analysis suggests that sequences with abundance of >100 copies contain no errors. Those with abundance of <100 could be potentially repaired. Validation of the error analysis and repair algorithm, however, is beyond the scope of this manuscript.

Positional analysis of amino acid abundances (Fig. 5) demonstrated that the distribution of amino acids in the top 150 se-

quences, present at copy number of >10,000, was different from that of the remaining library. Distribution of amino acids in sequences present at copy number <10,000 was similar to those in the overall library. Overall distribution of amino acids in peptides in the library was similar to those observed in earlier reports [17,21,29]. Library had abundant Ser/Thr in all positions. Abundance of Cys was low in all positions. N-terminus exhibited significant preference for some amino acids, presumably due to

proteolytic preference of the peptidase, which truncates leader peptide sequences following the displayed peptide [20,21].

Clustering analysis identified ten distinct sequence patterns in the top 150 fastest growing clones. Fig. 6 describes the clustering tree diagram and protein LOGO [30] display of the conserved sequence within each sub-sequence. Remarkably, a rare amino acid W appeared as a consensus amino acid in many sub-sequences, and it was present as the C-terminal amino acid in 50 out of 150 peptides. Our simple clustering analysis could be potentially replaced by more advanced software packages, such as Multiple Specificity Identifier (MUSI) [31], which was designed to identify distinct families of consensus sequence motifs within deep sequencing data. The analysis could potentially identify conserved peptide motifs emerging as the results of growth-induced selection.

7. Conclusions and future directions

Illumina sequencing, for the first time has uncovered a strong amplification bias to a small number of sequences. The scale at which this bias is visible is difficult to attain by other next-generation sequencing techniques. The reason for this bias remains unknown, but we strongly believe that the bias results from growth preferences of individual phage. It is unlikely to be the result of simple bias in PCR preparation; the latter bias is unlikely to give abundances of 10,000-fold. PCR also does not favor specific sequence but rather a class of sequences with specific melting point or specific GC-content [18,19]. The bias we observe is unlikely to be present in the naïve library, which should contain up to 10^9 clones according to the manufacturer (New England Biolabs). Indeed, sequencing of naïve (non-amplified) libraries demonstrated that there is little bias towards specific sequences in the library [14].

Deep sequencing of phage libraries also leaves a few open questions. One of them is general error analysis of random libraries. A growing body of literature confirms that a large number of errors are present in the Illumina results [28], but reliable identification of errors in random libraries is not trivial. The other unexplained observation is the dramatic abundance of reverse reads when compared to forward reads (Fig. 2). The preparation based on dsDNA should give equal number of forward and reverse strands; the reason for the observed bias towards reverse strands is unclear. It is unlikely that the reads are lost in the analysis because our analysis maps account for mutations and frame shifts of constant primer regions and, thus, can map up to >99% of reads. We hypothesize that hybridization to Illumina chip and on-chip sequencing might be biased towards one read (or one type of DNA sequence). On-chip sequencing is known to discriminate against specific classes of sequences and introduce specific errors (frame shifts, etc.) [28]. The analysis of sequence bias in different reads and comprehensive error analysis will be described in our subsequent manuscript. Overall, we foresee that Illumina sequencing and analysis similar to the one outlined in this manuscript will provide many advantages to the analysis of phage-display screens. Furthermore, analysis of the biological origin of sequences emerging from amplified libraries will enable identification of a mechanism that promotes or interferes with selection of useful binding sequences in phage display.

Acknowledgements

The authors thank Christian Daly and Claire Reardon at Harvard FAS sequencing center for helpful discussions. This work was supported by Alberta Glycomics Centre, University of Alberta Startup Funds, NSERC Discovery Grant (402511-2011), Canadian Founda-

tion for Innovation New Leaders Opportunity Grant, and DARPA InfoChemistry award (to R.D.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ymeth.2012.07.006>.

References

- [1] J.K. Scott, G.P. Smith, *Science* 249 (1990) 386–390.
- [2] G.P. Smith, V.A. Petrenko, *Chemical Reviews* 97 (1997) 391–410.
- [3] D.J. Rodi, R.W. Janes, H.J. Sanganeer, R.A. Holton, B.A. Wallace, L. Makowski, *Journal of Molecular Biology* 285 (1999) 197–203.
- [4] D.J. Rodi, G.E. Agoston, R. Manon, R. Lapcevic, S.J. Green, L. Makowski, *Combinatorial Chemistry & High Throughput Screening* 4 (2001) 553–572.
- [5] E. Dias-Neto, D.N. Nunes, R.J. Giordano, J. Sun, G.H. Botz, K. Yang, J.C. Setubal, R. Pasqualini, W. Arap, *Plos One* 4 (2009) e8338.
- [6] R. Di Niro, A.-M. Sulic, F. Mignone, S. D'Angelo, R. Bordon, M. Iacono, R. Marzari, T. Gaiotto, M. Lavric, A.R.M. Bradbury, L. Biancone, D. Zevin-Sonkin, G. De Bellis, C. Santoro, D. Sblattero, *Nucleic Acids Research* 38 (2010) e110.
- [7] E. Malini, E. Maurizio, S. Bembich, R. Sgarra, P. Edomi, G. Manfioletti, *Biochemistry* 50 (2011) 3462–3468.
- [8] A. Ernst, D. Gfeller, Z. Kan, S. Seshagiri, P.M. Kim, G.D. Bader, S.S. Sidhu, *Molecular Biosystems* 6 (2010) 1782–1790.
- [9] R. Tonikian, X. Xin, C.P. Torel, D. Gfeller, C. Landgraf, S. Panni, S. Paoluzi, L. Castagnoli, B. Currell, S. Seshagiri, H. Yu, B. Winsor, M. Vidal, M.B. Gerstein, G.D. Bader, R. Volkmer, G. Cesareni, D.G. Drubin, P.M. Kim, S.S. Sidhu, C. Boone, *Plos Biology* 7 (2009) e1000218.
- [10] A. Ernst, S.L. Sazinsky, S. Hui, B. Currell, M. Dharsee, S. Seshagiri, G.D. Bader, S.S. Sidhu, *Science Signaling* 2 (2009) ra50.
- [11] H. Zhang, A. Torkamani, T.M. Jones, D.I. Ruiz, J. Pons, R.A. Lerner, *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011) 13456–13461.
- [12] U. Ravn, F. Gueneau, L. Baerlocher, M. Osteras, M. Desmurs, P. Malinge, G. Magistrelli, L. Farinelli, M.H. Kosco-Vilbois, N. Fischer, *Nucleic Acids Research* 38 (2010) e193.
- [13] D.M. Fowler, C.L. Araya, S.J. Fleishman, E.H. Kellogg, J.J. Stephany, D. Baker, S. Fields, *Nature Methods* 7 (2010) 741–U108.
- [14] P.A.C. t Hoen, S.M.G. Jirka, B.R. ten Broeke, E.A. Schultes, B. Aguilera, K.H. Pang, H. Heemskerck, A. Aartsma-Rus, G.J. van Ommen, J.T. den Dunnen, *Analytical Biochemistry* 421 (2012) 622–631.
- [15] R. Derda, S. Musah, B.P. Orner, J.R. Klim, L.Y. Li, L.L. Kiessling, *Journal of the American Chemical Society* 132 (2010) 1289–1295.
- [16] R. Derda, S.K.Y. Tang, S.C. Li, S. Ng, W. Matochko, M.R. Jafari, *Molecules* 16 (2011) 1776–1803.
- [17] L. Makowski, A. Soares, *Bioinformatics* 19 (2003) 483–489.
- [18] B. Zimmermann, T. Gesell, D. Chen, C. Lorenz, R. Schroeder, *Plos One* 5 (2010).
- [19] W.H. Thiel, T. Bair, K.W. Thiel, J.P. Dassie, W.M. Rockey, C.A. Howell, X.Y.Y. Liu, A.J. Dupuy, L.Y. Huang, R. Owczarzy, M.A. Behlke, J.O. McNamara, P.H. Giangrande, *Nucleic Acid Therapeutics* 21 (2011) 253–263.
- [20] L. Makowski, in: V.A. Petrenko, G.P. Smith (Eds.), *Phage Nanobiotechnology*, RSC Publishing, 2011, pp. 33–54.
- [21] D.J. Rodi, A.S. Soares, L. Makowski, *Journal of Molecular Biology* 322 (2002) 1039–1052.
- [22] J. Huang, B. Ru, P. Zhu, F. Nie, J. Yang, X. Wang, P. Dai, H. Lin, F.-B. Guo, N. Rao, *Nucleic Acids Research* 40 (2012) D271–D277.
- [23] B. Ru, J. Huang, P. Dai, S. Li, Z. Xia, H. Ding, H. Lin, F.-B. Guo, X. Wang, *Molecules* 15 (2010) 8279–8288.
- [24] R. Derda, S.K.Y. Tang, G.M. Whitesides, *Angewandte Chemie International Edition* 49 (2010) 5301–5304.
- [25] W. Matochko, S. Ng, M.R. Jafari, R. J., S.K.Y. Tang, R. Derda, *Methods*, current (2012).
- [26] K.A. Noren, C.J. Noren, *Methods* 23 (2001) 169–178.
- [27] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, J.M. Boutell, J. Bryant, R.J. Carter, R.K. Cheetham, A.J. Cox, D.J. Ellis, M.R. Flatbush, N.A. Gormley, S.J. Humphray, L.J. Irving, M.S. Karbelashvili, S.M. Kirk, H. Li, X.H. Liu, K.S. Maisinger, L.J. Murray, B. Obradovic, T. Ost, M.L. Parkinson, M.R. Pratt, I.M.J. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M.T. Ross, A. Sabot, S.V. Sankar, A. Scally, G.P. Schroth, M.E. Smith, V.P. Smith, A. Spiridou, P.E. Torrance, S.S. Tzonev, E.H. Vermaas, K. Walter, X.L. Wu, L. Zhang, M.D. Alam, C. Anastasi, I.C. Aniebo, D.M.D. Bailey, I.R. Bancarz, S. Banerjee, S.G. Barbour, P.A. Baybayan, V.A. Benoit, K.F. Benson, C. Bevis, P.J. Black, A. Boodhun, J.S. Brennan, J.A. Bridgman, R.C. Brown, A.A. Brown, D.H. Buermann, A.A. Bundu, J.C. Burrows, N.P. Carter, N. Castillo, M.C.E. Catenazzi, S. Chang, R.N. Cooley, N.R. Crake, O.O. Dada, K.D. Diakoumakos, B. Dominguez-Fernandez, D.J. Earnshaw, U.C. Egbujor, D.W. Elmore, S.S. Etchin, M.R. Ewan, M. Fedurco, L.J. Fraser, K.V.F. Fajardo, W.S. Furey, D. George, K.J. Gietzen, C.P. Goddard, G.S. Golda, P.A. Granieri, D.E. Green, D.L. Gustafson, N.F. Hansen, K. Harnisch, C.D. Haudenschild, N.I. Heyer, M.M. Hims, J.T. Ho, A.M. Horgan, K. Hoshcler, S. Hurwitz, D.V. Ivanov, M.Q. Johnson, T. James, T.A.H. Jones, G.D. Kang, T.H.

- Kerelska, A.D. Kersey, I. Khrebtukova, A.P. Kindwall, Z. Kingsbury, P.I. Kokko-Gonzales, A. Kumar, M.A. Laurent, C.T. Lawley, S.E. Lee, X. Lee, A.K. Liao, J.A. Loch, M. Lok, S.J. Luo, R.M. Mammen, J.W. Martin, P.G. McCauley, P. McNitt, P. Mehta, K.W. Moon, J.W. Mullens, T. Newington, Z.M. Ning, B.L. Ng, S.M. Novo, M.J. O'Neill, M.A. Osborne, A. Osnowski, O. Ostadan, L.L. Paraschos, L. Pickering, A.C. Pike, A.C. Pike, D.C. Pinkard, D.P. Pliskin, J. Podhasky, V.J. Quijano, C. Raczy, V.H. Rae, S.R. Rawlings, A.C. Rodriguez, P.M. Roe, J. Rogers, M.C.R. Bacigalupo, N. Romanov, A. Romieu, R.K. Roth, N.J. Rourke, S.T. Ruediger, E. Rusman, R.M. Sanches-Kuiper, M.R. Schenker, J.M. Seoane, R.J. Shaw, M.K. Shiver, S.W. Short, N.L. Sizto, J.P. Sluis, M.A. Smith, J.E.S. Sohna, E.J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C.L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovskiy, S.M. Virk, S. Wakelin, G.C. Walcott, J.W. Wang, G.J. Worsley, J.Y. Yan, L. Yau, M. Zuerlein, J. Rogers, J.C. Mullikin, M.E. Hurles, N.J. McCooke, J.S. West, F.L. Oaks, P.L. Lundberg, D. Klenerman, R. Durbin, A.J. Smith, *Nature* 456 (2008) 53–59.
- [28] M. Kircher, P. Heyn, J. Kelso, *Bmc Genomics* 12 (2011) 382.
- [29] S.J. Lee, J.H. Lee, B.K. Kay, G. Dreyfuss, Y.K. Park, J.K. Kim, *Journal of Microbiology* 35 (1997) 347–353.
- [30] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, *Genome Research* 14 (2004) 1188–1190.
- [31] T. Kim, M.S. Tyndel, H. Huang, S.S. Sidhu, G.D. Bader, D. Gfeller, P.M. Kim, *Nucleic Acids Research* 40 (2012) e47.