Published in final edited form as:

Anal Biochem. 2011 April 1; 411(1): 106–115. doi:10.1016/j.ab.2010.11.035.

Design, construction, and validation of a modular library of sequence diversity standards for PCR

Paul D. Baum, Jennifer J. Young, Qianjun Zhang, Zeljka Kasakow, and Joseph M. McCune

Division of Experimental Medicine, University of California, San Francisco, Box 1234, San Francisco, CA, 94143-1234, Telephone (415) 826-9517, Fax (415) 206-8091

Abstract

Methods to measure the sequence diversity of PCR-amplified DNA lack standards for use as assay calibrators and controls. Here, we present a general and economical method for developing customizable DNA standards of known sequence diversity. Standards ranging from 1 to 25,000 sequences were generated by directional ligation of oligonucleotide “words” of standard length and GC content, and then amplified by PCR. The sequence accuracy and diversity of the library were validated using AmpliCot analysis (DNA hybridization kinetics) and Illumina sequencing. The library has the following features: (1) pools containing tens of thousands of sequences can be generated from the ligation of relatively few commercially-synthesized short oligonucleotides; (2) each sequence differs from all others in the library at a minimum of three nucleotide positions, permitting discrimination between different sequences by either sequencing or hybridization; (3) all sequences have identical length, GC content, and melting temperature; (4) the identity of each standard can be verified by restriction digestion; and (5) once made, the ends of the library may be cleaved and replaced with sequences to match any PCR primer pair. These standards should greatly improve the accuracy and reproducibility of sequence diversity measurements.

Keywords

sequence; diversity; PCR; standards; ligation; AmpliCot

INTRODUCTION

Many areas of biology use PCR to amplify complex nucleic acid templates. For example, studies of microbial ecology [1] and of viral quasispecies [2] require the sensitive and accurate amplification and enumeration of sequences, a difficult process prone to artifacts [3]. Other areas of investigation involve the construction of diverse libraries of sequences, followed by rounds of selection. Such schemes have been used to identify RNA or protein molecules with particular enzymatic or binding activities [4,5], compounds of interest from combinatorial chemistry libraries [6–8], or solutions to computationally difficult problems using DNA computing [9]. In these schemes, the diversity of the starting library is often assumed rather than measured, and selection process is only evaluated at the end of a long experiment. The effects of the selection conditions chosen or the number of iterations performed can be difficult to monitor at early stages of the experiment.

Correspondence should be addressed to P.D.B. (paul.baum@ucsf.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The accuracy and precision of such PCR-based experiments would be greatly enhanced by the inclusion of standards containing known numbers of sequences. Such standards are needed to establish the sensitivity and reproducibility of the amplification steps of the experiment, and to rule out PCR bottlenecks, contamination, or recombination. Sequence diversity standards could also help to detect artifacts generated in the steps after PCR, such as cloning, or the processing of samples for high throughput sequencing. In particular, equilibrium [10,11] or kinetic [12] hybridization-based methods for determining sequence diversity could use standards of known diversity, both to compensate for differences in reaction conditions from experiment to experiment and to adjust for any non-linearity in the assays. An ideal diversity standard would have end sequences that could be cleaved and customized to match any desired pair of PCR primers. This would allow a single library of standards to be used with multiple primer pairs, without requiring generation and validation of the standards each time. For example, standards for PCR-based assays measuring the diversity of T cell receptor VDJ gene rearrangements (our particular interest) would have ends tailored to be amplifiable by the primers for each particular V-J pair.

Existing methods for making DNA diversity standards have significant limitations for use with PCR amplification. The most commonly used standard, both for traditional Cot analysis and for next generation sequencing, is genomic DNA. Given biological selection pressures, genomic DNA has fixed DNA sequence diversity. Further, the representation of DNA sequences relative to one another is stable, albeit often skewed. However, genomic DNA cannot be amplified directly as a control for PCR experiments, and once it has been adapted for use with PCR, the advantages of stable, known sequence diversity are lost. Additionally, there are differences in genomic DNA preparations in terms of length distribution, ends, and salt concentration that may be a source of variability in some types of experiments [13].

Biological samples thought to be of high diversity may also be used as standards. For example, in a study of T cell receptor sequences, one might prepare amplicons from large numbers of cells pooled from different donors. This method, however, makes assumptions about the intrinsic diversity of the starting sample, limits the ability to make standards of very high diversity relative to the samples one hopes to measure, and may not be readily reproducible.

A third source of diverse sequences is the synthesis of oligonucleotides with degenerate positions [10,11]. This method can provide templates with the proper ends for PCR at very low cost. However, such standards have three important limitations. First, it is technically difficult to obtain an equal representation of nucleotides at each position; with large numbers of degenerate positions, the distribution may be highly skewed. Second, it is difficult to synthesize large amounts of long oligonucleotides free of failure sequences, limiting the size of PCR templates that can be directly synthesized. Finally, and most importantly, two members of a pool of degenerate oligonucleotide sequences can differ from each other at as few as one position. This makes it difficult to use such pools as standards for hybridization methods or for high throughput sequencing, where the ability to resolve one sequence from another requires them to be different at several nucleotide positions. Schütze and colleagues, for example, demonstrated that degenerate oligonucleotides annealed into heteroduplexes making them unsuitable for use as AmpliCot assay standards [11].

A final option is to synthesize separate oligonucleotides for each different sequence in the mix. While standard synthesis methods may be used to make and purify small numbers of oligonucleotides, these methods are too costly for making standards containing thousands or more sequences. Pools of oligonucleotides differing at multiple positions can be synthesized using split-pool synthesis or “word” phosphoramidites (containing multiple nucleotides), but

such services are not commercially available [14–16]. High throughput oligonucleotide synthesis using microfluidics or photolithography is currently still expensive and may result in contamination of the sequence pool with a high rate of failure sequences [17–19].

Here, we provide a general method for making a library of sequences of known, high diversity that can be used as standards for the measurement of diversity of PCR amplicons. The method uses modular construction that allows the diversity of the resulting sequences to be easily controlled. Combinatorial generation of diversity permits tens of thousands of sequences to be made from less than 100 short oligonucleotides. Additionally, the ends of a characterized library may be replaced with different sequences for use with different PCR primers. The use of different “word” sequences ensures that all sequences are of uniform GC content and length, resulting in uniform melting temperature and hybridization properties. Each word differs from every other word at a minimum of three nucleotide positions, therefore ensuring that the sequences made when words are concatenated will also differ from one other at a minimum of three nucleotide positions. This feature ensures that the library members can be distinguished by stringent hybridization and allows for the correction of single nucleotide misreads in high throughput sequencing. This simple, economical, and flexible method will permit the use of appropriate sequence diversity standards for calibrating and validating any type of nucleic acid diversity experiment. The method could also be adapted to make word-based libraries used for DNA computing and combinatorial chemistry methods.

MATERIALS AND METHODS

Oligonucleotide design and synthesis

Word sequences, containing six nucleotides of fixed GC content, were selected by computer so that each differed from every other sequence in the set at a minimum of three positions (software available on request). We note that more sophisticated heuristics for DNA word selection have been developed [20,21], but we did not use them because they have not, to our knowledge, been fully experimentally validated. The selected words were used to design a set of short oligonucleotides from which the library was constructed. See Figure 1 for design and Table 1 for sequences of the LP and RP (Left and Right PCR) primers as well as of the L (Left), R (Right), T (Top), and B (Bottom) word oligonucleotides. Oligonucleotides (Operon and Integrated DNA Technologies) were chemically 5' phosphorylated (T and R oligos only), purified with HPSF or HPLC chromatography, and normalized for concentration. Oligonucleotide pairs were annealed in a 10 mM TrisCl pH 8, 1 mM EDTA 100 mM NaCl buffer over two hrs. T and B oligonucleotide pairs were annealed in a 65°C to 4°C gradient and library end adapters were annealed in a 80°C to 25°C gradient.

PCR amplification

All amplifications of the library took place in 25 μ L reactions using 500 nM PCR primers and 0.5 units hot start Phusion polymerase with buffer HF (NEB). Thermocycling consisted of a 98°C hot start for 30 secs, followed by cycles of 98°C for 10 sec, 62°C for 15 sec, and 72°C for 15 sec. Cycling concluded with a 5 min incubation at 72°C. Many protocol steps (e.g., analytical and preparative restriction digestion, melting curves, and AmpliCot analysis) required the PCR product to be in homoduplex form. This was achieved by amplification for 15 cycles, followed by a 1:10 dilution of the first-round PCR product into a fresh 25 μ L PCR reaction that was amplified for three more cycles. For real-time PCR measurements on an Opticon machine (Bio-Rad), we used SYBR Green mastermix (MC Lab) with 500 nM LP and RP primers and cycling parameters of a 95°C hot start for 5 min, followed by cycles of 95°C for 15 sec and 60°C for 30 sec.

Combinatorial library synthesis

Equimolar mixtures of L, R, and annealed T/B oligonucleotides (10 μ mol each) were combined according to the scheme in Table 2. T4 ligase (NEB) was used to ligate the single-stranded L and R oligonucleotides to the T oligonucleotides. The B oligonucleotides provided a double-stranded region required for the T4 ligase activity. Reactions containing 400 Cohesive End Units ligase in a total volume of 10 μ L were incubated overnight at 4°C. Ligation products were diluted 1:25,000 and amplified for 22 cycles with LP and RP PCR primers. PCR products were run on denaturing 10% polyacrylamide minigels (Invitrogen) and stained with ethidium bromide. Single bands of the expected 66 nucleotide length were excised from the gel, and DNA was eluted by passive diffusion overnight into 500 μ L TE buffer at 4°C and then stored at -20°C. Generally, a 1:1000 dilution of 1 μ L of this eluate sufficed as a PCR template for amplifying these standards, but larger volumes were used for preparation of library stocks with modified ends. A detailed protocol and directions for requesting aliquots of the library are at <http://amplicot.ucsf.edu>.

Modification of library end sequences

To minimize the number of amplification cycles and to avoid a loss of diversity due to bottlenecks, eluate (8 μ L) from each gel purified standard was used as a PCR template. To remove the left end, 25 μ L of unpurified homoduplex PCR product were digested with 20 units BsmBI (NEB) in a 50 μ L reaction for two hrs at 42°C. (Digestion of library ends was carried out at temperatures below the maximal activity of the thermophilic enzymes used to minimize melting of short library fragments.) Double-stranded V gene adapters (100 nM) were ligated to 2 L of the digest in a 10 μ L reaction with 400 Cohesive End Units of T4 ligase at 25°C for 1 hr. The ligation products were diluted 1:250 and then amplified with V and RP primers, followed by 1:10 dilution into a fresh reaction for three cycles to make homoduplex products (as above). To remove the right ends, 25 μ L of this unpurified homoduplex PCR product were then digested with 20 units BsaI in a 50 μ L reaction for two hrs at 42°C. 2 μ L of the digest were ligated to double stranded J gene adapters (100 nM), as above. The ligation products were diluted 1:250, amplified with V and J primers, and then purified by polyacrylamide gel electrophoresis, as described above.

Analytical restriction digestion and melting curves

Homoduplex PCR products were digested without purification in PCR buffer with 10 units of restriction enzymes: AluI, HindIII, RsaI, or Sau3aI (NEB). For melting curves, AmpliCot buffer concentrate was added to undigested or digested PCR products to achieve a final concentration of 20 mM MOPS, pH 7.5, 10 mM EDTA, 250 mM NaCl, 0.03% Brij-700, and 5x SYBR Green dye. Melting curves were measured over a 70–95°C gradient in an Opticon real time PCR machine (Bio-Rad).

AmpliCot (DNA hybridization kinetics) analysis

100 μ L of homoduplex PCR product were treated with 20 units of Exonuclease I (NEB) for 1 hr at 37°C. PCR products were then mixed with 25 μ L of AmpliCot buffer (as above). 10 μ L of each PCR product were used to measure a melting curve. Annealing was then carried out in 50 μ L reactions, as described [12] at a temperature 3.5°C below the melting temperature. $Cot_{1/2}$ values were calculated by multiplying the concentration (in arbitrary fluorescence units) by the time in seconds required for 50% of the sample to anneal.

Sequencing of library

Homoduplex PCR products were cloned into the pCR2.1 T vector (Invitrogen). Transformants were cultured, prepped with the Qiaprep 96 kit (Qiagen), and used for Sanger sequencing. Illumina sequencing required modifying the ends of the library, using the

strategy outlined above. The sequences of oligonucleotides used are shown in Table 3. Oligonucleotides were annealed, as described above, to form thirteen sets of LT and LB (Left Top and Left Bottom) Gex1 adapter pairs, containing unique sequence tags and a single Gex2 adapter pair. All oligonucleotides used for preparing the sequencing samples were PAGE-purified and the PCR primers were synthesized with phosphorothioate linkages at their 3' ends to protect them from the exonuclease activity of the polymerase [22]. Using the procedure detailed above, the left end of each of the 13 diversity standards was ligated to a unique Gex1 adapter pair, followed by amplification with Gex1 and RP primers, ligation of the right end of the library to the Gex2 adapter sequence, and amplification with Gex1 and Gex2 primers. The final PCR product was purified with a denaturing polyacrylamide gel and eluted as above. DNA was ethanol precipitated and concentrations checked by spectrophotometry (Nanodrop). A mixture (weighted to include more of the high diversity standards) was sequenced for 42 cycles on an Illumina GA-II, using the NlaIII RNaseq primer (Harry Gao, City of Hope, Loma Linda, CA). Sequences filtered for Phred scores >5 at all positions were analyzed with Awk scripts to measure relative abundances of words and word combinations. Sequences filtered for Phred scores >20 at all positions were analyzed to estimate point and frameshift mutation rates.

RESULTS

To create a modular set of PCR diversity standards at low cost, we used a combinatorial word strategy [23,24]. Oligonucleotides containing six-nucleotide words were joined by ligation and amplified by PCR. An overview of the method is provided in Figure 1 and oligonucleotide sequences are shown in Table 1. The total length of the ligation products was a compromise between three goals: maximizing sequence diversity, allowing adequate length for annealing steps in library construction, and keeping the total length of the standards similar to that of the sequences being measured in the assay. Words were selected to be of equal GC content and length and to differ from each other at a minimum of three positions. To maximize fidelity and equal representation, oligonucleotides were synthesized with chemical phosphorylation, chromatographically purified, and then normalized for concentration. Ligation sites contained conserved overhang sequences to drive directional ligation and to minimize sequence-dependent bias in ligation efficiency or word association. The choice of sequences and their method of combination were further designed to create sites for diagnostic restriction enzyme digestion. Words were combined in defined pools to create different amounts of diversity at each word position. The smaller pools were non-overlapping so that the identities of most of the resulting standards could be verified through restriction digestion or sequencing of a single clone (Table 2).

Oligonucleotide ligation products were amplified with high fidelity Phusion polymerase to minimize PCR-related errors and PCR products were purified by gel electrophoresis to avoid unwanted side products. Each of these steps was found to be necessary (data not shown). After validation of the core library, type IIs restriction sites were used to create sticky ends that allowed ligation of adapter sequences for primer pairs of T cell receptor amplicons and for high throughput sequencing. Sequential cleavage and ligation steps for each end of the library greatly improved product yield and purity when compared to simultaneous replacement of both ends (i.e., two two-part ligations rather than a single three-part ligation).

The oligonucleotide standards were designed to have uniform homoduplex melting temperatures (by virtue of identical length and GC content) and also to have a wide separation between heteroduplexes and homoduplexes (due to word-based design). These properties are critical for hybridization-based assays and were confirmed by melting curve analysis using saturating concentrations of SYBR Green dye, as shown in Figure 2. Panel A,

obtained from standards starting in homoduplex form, demonstrates that the sequences did have similar melting temperatures, as designed. Panel B shows melting curves obtained from standards that were allowed to anneal at low stringency before the melting curves were generated. Peaks consistent with 0, 1, 2, and 3 word mismatches were detected in standards that contained variable words in at least those numbers of positions.

The oligonucleotide words at each position were designed so that some of the words created restriction sites, while others did not. Words were pooled at a given position so that all words had the same cleavage properties. Combinations of word pools at all three positions were then chosen so that restriction digestion of a specific diversity standard would yield an unambiguous cleavage pattern that could be used to confirm its identity, as outlined in Table 2. This was verified by restriction digestion of the PCR products followed by gel electrophoresis or melting curves. Because creation of the restriction sites required the successful ligation of different oligonucleotide building blocks, the fact that the library sequences could be cleaved as designed also validated the success of the ligation steps. The lower diversity standards were also composed of mutually exclusive word sets, so that the identity of a particular standard could be verified with a small number of sequencing reads.

We validated the actual diversity of the constructed standards using DNA hybridization kinetics, Sanger sequencing, and Illumina sequencing. The gap between the 0 and 1 word mismatches, shown in Figure 2B, suggested that an annealing temperature could be chosen that would selectively allow only perfectly matched duplexes to form, thereby permitting DNA hybridization kinetics to be used to measure sequence diversity. DNA hybridization kinetics showed a linear relationship between the combinatorial sequence diversity of the template and the $Cot_{1/2}$ value (concentration-time product required for 50% of the sequences to re-anneal), as measured with the AmpliCot assay (Figure 3). To further test our combinatorial model, the $Cot_{1/2}$ values were measured for standards that contained different word combinations that should equal the same amount of total diversity. As expected, combinations that yielded the same total diversity manifested similar $Cot_{1/2}$ values in the AmpliCot assay (Figure 4).

Finally, the diversity standards were analyzed by cloning and Sanger sequencing, as well as by Illumina sequencing. To measure mutation rates in our sequences, Illumina sequences with Phred scores >20 at each position were studied. 94% of sequences were perfect, containing the words and spacer sequences that were designed. Approximately 1% of sequences contained frameshift errors and 5% contained point mutations. Some but not all of these mutations may be due to the Illumina sequencing and not to the templates themselves (the error rate for the PhiX genomic DNA control was 0.4%/base). Errors were also detected in 5% (4 of 84) Sanger sequenced plasmid clones derived from the library.

By analogy to molecular cloning, we expected most errors to occur at the ligation sites used to construct the library. To our surprise, the point mutation rate as measured by Illumina sequencing was higher in the “word” nucleotides that differed between sequences than in the “spacer” nucleotides that were shared among all sequences on the plate (Figure 5). The sequence errors are unlikely to be due to recombination during preparation of the library, because they were seen in collections of sequences that were independently prepared and contained only a single clone. There are two potential explanations for this finding. First, the ligation step may have selected against oligonucleotides containing errors at the hybridized ends. Ligation is well known to require perfect base pairing, and a hybridization-ligation strategy has been used to reduce the error rate in genes synthesized from pools of imperfect oligonucleotides [24,25]. Second, it is possible that on a densely packed Illumina flow cell, the presence of a vast majority of sequences displaying a single base reduces the sensitivity for minority nucleotides at that position. While we cannot exclude the former explanation,

we favor the latter because the mutation rates at the common positions of the library are actually lower than the mutation rates measured for the PhiX control, consistent with suppression of variant base calling. Additionally, the same effect of higher error rates in variable regions was seen at the barcode sequences at the beginning of the reads, sequences which were all generated by the double-stranded oligonucleotide adapters used to prepare the library for Illumina sequencing.

We then turned our attention to the relative representation of the words used in the library, which was expected to be uniform based on the construction methods. Figure 6 shows the relative use of words at each position. To increase our sequence coverage, we counted over 19 million sequences with Phred scores >5 at each position. Less than 1% of the tagged sequences contained words that were not supposed to be in their ligation mixture, suggesting that minimal recombination occurred between similar amplicons during the multiplexed cluster generation process for Illumina sequencing. However, there was an unexpected skewing of the use of the words in the left position, a bias that was consistently observed in independent ligation mixtures (Figure 6A). This skewing was presumably due to preferential ligation or amplification, which we are unable to explain on the basis of oligonucleotide quality or sequence. The combinations of these various words were consequently unequally represented (Figure 6B).

Despite this skewed representation of sequences, every expected sequence combination was detected for the lower diversity standards, A-G, that had over 1000 reads per expected sequence combination. For the most complex standards, H-M, which were less sampled relative to their diversity, we were still able to find at least 95% of the expected word combinations. The abundance of various word combinations correlated well with the product of the abundances of the words used to make them (Spearman's r of 0.97, Figure 6C), suggesting that there was no particular bias in which word combinations could be made. The high diversity of the word combination products suggested that the full diversity of the starting library was conserved through the ligation, amplification, and gel purification steps performed to change the end sequences of the library. Simulations of the effects of this observed amount of skewing would have on AmpliCot annealing curves (not shown) indicate that the skewed standards would have an apparent 50% decrease in $Cot_{1/2}$ measurements as compared to standards that had precisely equal representation of all component sequences. Given that the unknowns to be measured with the AmpliCot assay are likely to also have unequal sequence abundances, however, the error may be less than this estimate in practice.

To illustrate how these standards may be run alongside experimental samples in order to monitor the quality of a PCR amplification, we made serial dilutions of two standards containing different sequence diversity. These dilutions were used as templates for PCR and then the sequence diversity of these PCR products was measured with AmpliCot (Figure 7). At low levels of dilution, the standards showed their expected diversity. With increasing levels of dilution, threshold points were reached where the measured diversity began to decrease. The more complex standard reached its threshold at relatively lower level of dilution. At the highest levels of dilution, both standards showed similar decreased levels of diversity. These results suggest that there was a bottleneck effect at high dilutions caused by too few template molecules in the PCR reaction so that the full diversity of the template was not captured. In experiments measuring the diversity of complex templates, sequence diversity standards can be used in this fashion to determine the minimum amount of template that must be used in the PCR reaction to be able to measure a given amount of sequence diversity.

DISCUSSION

We describe here a general method for creating amplicon standards of known sequence diversity that can be used as a control in experiments measuring the sequence diversity of PCR products, much as abundance standards are used for quantitative PCR experiments. We present a novel strategy of ligating varying numbers of small oligonucleotide modules to create sequences of known sequence diversity. The protocol can be completed at low cost with standard laboratory equipment. The ends of the sequences can be modified to allow the amplicons to be used with any PCR primer pair. The library contains sequences of equal GC content and length, which therefore have a uniform melting temperature, as was experimentally confirmed. Each member of the library differs from each other member at a minimum of three nucleotide positions. This ensures that matched and mismatched sequences will be distinguishable by stringent hybridization (because of well-separated melting temperatures) and allows for the distinction of different sequences by sequencing (functioning as an error correcting code to tolerate single position mutations or sequencing errors). Finally, the identity of standards made with this library may be verified relatively easily, either through restriction digestion after PCR amplification or through sequencing of a single clone.

We found that the ligated sequences formed as predicted. Since the sequences were derived from a limited number of short building blocks that could be purified before use, the error rate in the final product was minimized. Reports of high throughput, long oligonucleotide synthesis indicate that the majority of sequences made with these other methods contain errors [17–19], which would be problematic for use as sequence diversity standards.

While our method provides many advantages over existing approaches, several limitations remain. First, given the fact that the component words were not equally likely to be used and the multiplier effect of the combinatorial library, the final sequences were skewed in abundance. This problem may be ameliorated in the future by empirical selection of oligonucleotides that are equally represented in test ligation experiments. For applications in which it is critical to obtain a more even representation of the sequences in the set, it might be possible to add a normalization step consisting of selection by hybridization to limiting amounts of oligonucleotides complementary to the words used. Second, because of overhang nucleotides required for ligation to construct this library, the maximum diversity that a library can attain for a given length is less than a library constructed through, for example, split-pool or direct synthesis. Third, the requirement for gel purification is time consuming and creates a risk for laboratory contamination with the sequences.

Our validation tests of the library also provide information about the analysis methods used. First, the hybridization kinetics data provide the first direct evidence that the AmpliCot assay can measure sequence diversity to the level of tens of thousands, at least of sequences that differ from each other at a minimum of three positions. Previous tests with individually synthesized oligonucleotide standards were only able to show that the assay was able to detect 96 sequences. Second, the sequencing analysis provides evidence that recombination of related sequences is minimal during the preparation of Illumina sequencing samples [27]. It also suggests that, at high cluster densities, Illumina sequencing may slightly underestimate the true prevalence of minority sequences.

The method presented here allows, at minimal cost, the construction of highly diverse sequence standards with ends that match any desired PCR primer sequences. For experiments that measure the sequence diversity of PCR products, whether through sequencing or other methods, these standards provide a way to monitor the PCR process, to make sure that enough template is used to avoid bottleneck effects, and to make sure that the

amplification and sequencing methods are not artificially increasing the amount of diversity in the sample. While informatics methods may be used to adjust for PCR or sequencing based artifacts, constructed templates of known diversity are a critical tool for validating these assays [9]. For AmpliCot measurements of T cell receptor diversity, typical experiments measuring the diversity of rearrangements using a selected V and J gene pair contain up to 10,000 sequences. The standards described here will now permit AmpliCot measurements to be interpolated, rather than extrapolated, greatly improving the accuracy of the assay. In addition to AmpliCot, these standards may find use for any sequencing or hybridization-based assay where individual sequences must differ from each other at multiple positions to be resolved. Finally, our methods for constructing the standards may be a cost-effective means for generating libraries for DNA computing or combinatorial chemistry.

Acknowledgments

FUNDING

This work was funded by grants from the National Institutes of Health [K23 AI 073100 to P.D.B., and R37 AI40312 and U01 AI43864 to J.M.M.] and from the California HIV/AIDS Research Program [an IDEA Grant to P.D.B.]. J.M.M. is a recipient of the NIH Director's Pioneer Award Program, part of the NIH Roadmap for Medical Research, through grant DPI OD00329.

We thank Teri Liegler, Jean-Francois Poulin, Dennis Hartigan-O'Connor, and McCune lab members for discussions and Margaret Lowe and Adam Lauring for comments on the manuscript. We thank Harry Gao and Nick Nelson for discussions of Illumina sequencing data, and Saunak Sen for statistical expertise.

References

1. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the human intestinal microbial flora. *Science*. 2005; 308:1635–1638. [PubMed: 15831718]
2. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res*. 2007; 17:1195–1201. [PubMed: 17600086]
3. Reeder J, Knight R. The 'rare biosphere': a reality check. *Nature Methods*. 2009; 6:636–7. [PubMed: 19718016]
4. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. *Nature*. 1990; 346:818–822. [PubMed: 1697402]
5. Scott JK, Smith GP. Searching for peptide ligands with an epitope library. *Science*. 1990; 249:386–390. [PubMed: 1696028]
6. Needels MC, Jones DG, Tate EH, Heinkel GL, Kochersperger LM, Dower WJ, Barrett RW, Gallop MA. Generation and screening of an oligonucleotide-encoded synthetic peptide library. *Proc Natl Acad Sci U S A*. 1993; 90:10700–10704. [PubMed: 7504279]
7. Halpin DR, Harbury PB. DNA display I. sequence-encoded routing of DNA populations. *PLoS Biol*. 2004; 2:E173. [PubMed: 15221027]
8. Clark MA, Acharya RA, Arico-Muendel CC, Belyanskaya SL, Benjamin DR, Carlson NR, Centrella PA, Chiu CH, Creaser SP, Cuzzo JW, Davie CP, Ding Y, Franklin GJ, Franzen KD, Geftter ML, Hale SP, Hansen NJ, Israel DI, Jiang J, Kavarana MJ, Kelley MS, Kollmann CS, Li F, Lind K, Mataruse S, Medeiros PF, Messer JA, Myers P, O'Keefe H, Oliff MC, Rise CE, Satz AL, Skinner SR, Svendsen JL, Tang L, van Vloten K, Wagner RW, Yao G, Zhao B, Morgan BA. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat Chem Biol*. 2009; 5:647–654. [PubMed: 19648931]
9. Braich RS, Chelyapov N, Johnson C, Rothmund PW, Adleman L. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*. 2002; 296:499–502. [PubMed: 11896237]
10. Ogle BM, Cascalho M, Joao CM, Taylor WR, West LJ, Platt JL. Direct measurement of lymphocyte receptor diversity. *Nucleic Acids Res*. 2003; 31:e139. [PubMed: 14602932]

11. Schütze T, Arndt PF, Menger M, Wochner A, Vingron M, Erdmann VA, Lehrach H, Kaps C, Glökler J. A calibrated diversity assay for nucleic acid libraries using DiStRO--a diversity standard of random oligonucleotides. *Nucleic Acids Res.* 2010; 38:e23. [PubMed: 19965765]
12. Baum PD, McCune JM. Direct measurement of T cell receptor repertoire diversity with AmpliCot. *Nature Methods.* 2006; 3:895–901. [PubMed: 17060913]
13. Young, BD.; Anderson, MLM. Quantitative analysis of solution hybridisation. In: Hames, BD.; Higgins, SJ., editors. *Nucleic acid hybridisation: a practical approach.* IRL Press; Oxford: 1985. p. 47-71.
14. Virnekas B, Ge L, Pluckthun A, Schneider KC, Wellnhofer G, Moroney SE. Trinucleotide phosphoramidites: Ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. *Nucleic Acids Res.* 1994; 22:5600–7. [PubMed: 7838712]
15. Sondek J, Shortle D. A general strategy for random insertion and substitution mutagenesis: Substoichiometric coupling of trinucleotide phosphoramidites. *Proc Natl Acad Sci U S A.* 1992; 89:3581–5. [PubMed: 1565654]
16. Tabuchi I, Soramoto S, Ueno S, Husimi Y. Multi-line split DNA synthesis: A novel combinatorial method to make high quality peptide libraries. *BMC Biotechnol.* 2004; 4:19. [PubMed: 15341664]
17. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature.* 2004; 432:1050–4. [PubMed: 15616567]
18. Cleary MA, Kilian K, Wang Y, Bradshaw J, Cavet G, Ge W, Kulkarni A, Paddison PJ, Chang K, Sheth N, Leproust E, Coffey EM, Burchard J, McCombie WR, Linsley P, Hannon GJ. Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat Methods.* 2004; 1:241–8. [PubMed: 15782200]
19. Lee CC, Snyder TM, Quake SR. A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res.* 2010; 38:2514–21. [PubMed: 20176572]
20. Tulpan D, Andronescu M, Chang SB, Shortreed MR, Condon A, Hoos HH, Smith LM. Thermodynamically based DNA strand design. *Nucleic Acids Res.* 2005; 33:4951–64. [PubMed: 16145053]
21. Shortreed MR, Chang SB, Hong D, Phillips M, Campion B, Tulpan DC, Andronescu M, Condon A, Hoos HH, Smith LM. A thermodynamic approach to designing structure-free combinatorial DNA word sets. *Nucleic Acids Res.* 2005; 33:4965–77. [PubMed: 16284197]
22. de Noronha CM, Mullins JL. Amplimers with 3'-terminal phosphorothioate linkages resist degradation by vent polymerase and reduce Taq polymerase mispriming. *PCR Methods Appl.* 1992; 2:131–6. [PubMed: 1335814]
23. Frutos AG, Liu Q, Thiel AJ, Sanner AM, Condon AE, Smith LM, Corn RM. Demonstration of a word design strategy for DNA computing on surfaces. *Nucleic Acids Res.* 1997; 25:4748–57. [PubMed: 9441280]
24. Brenner S, Williams SR, Vermaas E, Storck T, Moon K, McCollum C, Mao JJ, Luo S, Kirchner JJ, Eletr S, DuBridge RB, Burcham T, Albrecht G. In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differently expressed cDNAs. *Proc Natl Acad Sci U S A.* 2000; 97:1665–70. [PubMed: 10677516]
25. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. *Science.* 1988; 241:1077–80. [PubMed: 3413476]
26. Smith HO, Hutchison CA 3rd, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: PhiX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci U S A.* 2003; 100:15440–5. [PubMed: 14657399]
27. Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Res.* 1990; 18:1687–91. [PubMed: 2186361]

Figure 1. Design of the diversity standards

A. Diversity standards were made by combining three groups of oligonucleotides: L and R single-stranded oligonucleotides, and double-stranded T and B oligonucleotide pairs. The library contained 6-nucleotide-long “words” in three positions indicated by highlighted type. The ligated oligonucleotides were amplified with primers LP and RP.

B. The presence or absence of a nucleotide at the end of each word was designed to determine whether a restriction site (boxed) was created or not in the sequence. Additionally, one of the left position words was designed to contain a HindIII restriction site.

C. Restriction sites for the type II enzymes, BsmBI and BsaI, were placed in the flanking sequences used for the original library construction, to allow the ends of the library to be cleaved and then changed as necessary.

D. Ends of the library may be replaced by ligating double-stranded adapters to the library. In this case, mouse T cell receptor V and J segment sequences are added to make a standard that can be amplified alongside T cell receptor sequences.

Figure 2. Melting curves of the diversity standards

All melting curves show the negative first derivative of the intensity of fluorescence (on the y axis) as the temperature of the sample is increased, as shown on the x axis. Standards containing sequence diversity at 0, 2, or 3 word positions are shown. (See Table 2B for description of combinations.)

A. Melting curves prepared starting from homoduplex DNA, at the end of DNA synthesis. A single sharp homoduplex peak is seen at the same temperature for all standards.

B. Melting curves starting from DNA that had been melted and allowed to re-anneal at low stringency. In addition to the homoduplex peak, several lower temperature peaks are seen consistent with word mismatches at 1, 2, or 3 positions.

Figure 3. AmpliCot validation of the diversity standards

Standards were prepared containing different amounts of theoretical diversity, as outlined in Table 2. Cot_{1/2} values were measured for these standards using the AmpliCot protocol; these are displayed in arbitrary units.

Figure 4. Validation of the combinatorial diversity principle of the library

Sets of standards containing the same amounts of theoretical diversity were prepared by using different numbers of sequences at different word positions. Samples with identical theoretical diversity are grouped together. Cot_{1/2} values were measured for these standards using the AmpliCot protocol; these are displayed in arbitrary units.

Figure 5. Varying error rate of Solexa sequencing

The single sequence standard (described in Table 2), which contains only one word in each position, should have a uniform sequence and provides a chance to measure the error rate at each nucleotide in the sequence. Reads commenced at the beginning of the five nucleotide barcode contained in the Solexa sequencing adapter and proceeded through the three “words” of the library. Variable regions of the entire standard library (the barcode and the three 6-nucleotide “words”) are indicated in red while nucleotides that should be uniform amongst all sequences are indicated in black. Error rates are indicated on the y axis. To avoid counting possible contaminant sequences, only sequences containing a single error relative to the predicted standard A sequence were included.

Figure 6. Representation of words and word combinations in the library

A. Relative representation of component words in standard M of the library (described in Table 2), which should contain all possible words, ideally at equal frequency. Similar results were seen in other library standards. Words are arranged from least represented to most represented. Word representation is shown in $\log_2$ transformation; the expected representation if all words were equally present would be at 0, indicated by the horizontal line. Representation of words at the left position is more skewed than the representation of words at the middle and right positions.

B. Relative representation of word combinations in several standards from the library. $\log_2$ transformation of representation of words at the 10th through the 90th percentile is shown, as

ranked by relative abundance. Again, the expected representation if all word combinations were equally present would be at 0, indicated by the horizontal line. The 64 sequence standard has relatively low skewing because it only contained a single word at the left position.

C. Representation of word combinations in a total of 2.1 million sequence reads of standard K of the library (8750 expected sequences) is shown on the y axis, compared to the representation predicted by multiplying the relative representations of the three component words of each combination, as shown on the x axis.

Figure 7. Diversity standards define a bottleneck threshold

Two standards of 1568 (■) and 25000 (▼) sequences were serially diluted, PCR-amplified, and then analyzed by AmpliCot. The x axis shows the quantitative PCR threshold cycle value, an indication of the amount of template in the reaction. The measured amount of diversity of the PCR product amplified from that template dilution is shown on the y axis.

Oligonucleotide Sequences

Table 1

Sequences are given for the PCR primers (LP and RP) and the Left (L), Right (R), Top middle (T) and Bottom middle (B) oligonucleotide components ligated together to make the diversity standards. The T and R oligonucleotides were ordered chemically phosphorylated to allow for uniform ligation efficiencies. The six nucleotide word regions are in bold. The L1-25 oligonucleotides in red generate ligation products cleavable with AluI; L40 in blue is cleavable with both AluI and HinDIII. The T1-16 and B1-16 oligonucleotides in red generate ligation products cleavable with RsaI. The R1-16 oligonucleotides in red are cleavable with Sau3AI.

PCR Primers	
LP	CCTGCAGGCGTCTCGATAC
RP	GGTCGACCGGTCTCTGAT
Word oligonucleotides	
L1	CCTGCAGGCGTCTCGATACCCAAAAGCT
L2	CCTGCAGGCGTCTCGATACGACAAAGCT
L3	CCTGCAGGCGTCTCGATACAGGAAAGCT
L4	CCTGCAGGCGTCTCGATACTGACAAGCT
L5	CCTGCAGGCGTCTCGATACATCCAAGCT
L6	CCTGCAGGCGTCTCGATACCATCAAGCT
L7	CCTGCAGGCGTCTCGATACTCCTAAGCT
L8	CCTGCAGGCGTCTCGATACGGTTAAGCT
L9	CCTGCAGGCGTCTCGATACCTGTAAGCT
L10	CCTGCAGGCGTCTCGATACGTAGAAGCT
L11	CCTGCAGGCGTCTCGATACACTGAAGCT
L12	CCTGCAGGCGTCTCGATACTAGGAAGCT
L13	CCTGCAGGCGTCTCGATACTTCACAGCT
L14	CCTGCAGGCGTCTCGATACAAACCAGCT
L15	CCTGCAGGCGTCTCGATACGGAATAGCT
L16	CCTGCAGGCGTCTCGATACACCATAGCT
L17	CCTGCAGGCGTCTCGATACCAGATAGCT
L18	CCTGCAGGCGTCTCGATACCTACTAGCT
L19	CCTGCAGGCGTCTCGATACTACCTAGCT
L20	CCTGCAGGCGTCTCGATACAGTCTAGCT
L21	CCTGCAGGCGTCTCGATACGTCTTAGCT
L22	CCTGCAGGCGTCTCGATACCCCTTAGCT
L23	CCTGCAGGCGTCTCGATACTGGTTAGCT
L24	CCTGCAGGCGTCTCGATACTCAGTAGCT
L25	CCTGCAGGCGTCTCGATACGATGTAGCT
L26	CCTGCAGGCGTCTCGATACTGCAATGCT
L27	CCTGCAGGCGTCTCGATACGCTAATGCT
L28	CCTGCAGGCGTCTCGATACACATGCT
L29	CCTGCAGGCGTCTCGATACTTGCATGCT
L30	CCTGCAGGCGTCTCGATACCGATATGCT
L31	CCTGCAGGCGTCTCGATACAACGATGCT

L32 CCTGCAGGCGTCTCGATACCTAACTGCT
L33 CCTGCAGGCGTCTCGATACACTTGTGCT
L34 CCTGCAGGCGTCTCGATACTATCCTGCT
L35 CCTGCAGGCGTCTCGATACTCATCTGCT
L36 CCTGCAGGCGTCTCGATACGTTTCTGCT
L37 CCTGCAGGCGTCTCGATACTACTGTGCT
L38 CCTGCAGGCGTCTCGATACCAAGTTGCT
L39 CCTGCAGGCGTCTCGATACTTCGTTGCT
L40 CCTGCAGGCGTCTCGATACAAGCTTGCT
R1 [phos]CACCCAAAGATCAGGAGACCGGTCGACC
R2 [phos]CACGACAAGATCAGGAGACCGGTCGACC
R3 [phos]CACAGGAAGATCAGGAGACCGGTCGACC
R4 [phos]CACTGACAGATCAGGAGACCGGTCGACC
R5 [phos]CACATCCAGATCAGGAGACCGGTCGACC
R6 [phos]CACCATCAGATCAGGAGACCGGTCGACC
R7 [phos]CACTCCTAGATCAGGAGACCGGTCGACC
R8 [phos]CACGGTTAGATCAGGAGACCGGTCGACC
R9 [phos]CACCTGTAGATCAGGAGACCGGTCGACC
R10 [phos]CACGTAGAGATCAGGAGACCGGTCGACC
R11 [phos]CACACTGAGATCAGGAGACCGGTCGACC
R12 [phos]CACTAGGAGATCAGGAGACCGGTCGACC
R13 [phos]CACTTCACGATCAGGAGACCGGTCGACC
R14 [phos]CACAACCGATCAGGAGACCGGTCGACC
R15 [phos]CACGGAATGATCAGGAGACCGGTCGACC
R16 [phos]CACACCATGATCAGGAGACCGGTCGACC
R17 [phos]CACCAAGATGATCAGGAGACCGGTCGACC
R18 [phos]CACTGCAACATCAGGAGACCGGTCGACC
R19 [phos]CACGCTAACATCAGGAGACCGGTCGACC
R20 [phos]CACACACATCAGGAGACCGGTCGACC
R21 [phos]CACTTGACATCAGGAGACCGGTCGACC
R22 [phos]CACCGATACATCAGGAGACCGGTCGACC
R23 [phos]CACAACGACATCAGGAGACCGGTCGACC
R24 [phos]CACCTAACCATCAGGAGACCGGTCGACC
R25 [phos]CACACTTGCATCAGGAGACCGGTCGACC
T1 [phos]ACGTAGAGTA
T2 [phos]ACACTGAGTA
T3 [phos]ACTAGGAGTA
T4 [phos]ACTTCACGTA
T5 [phos]ACAAACCGTA
T6 [phos]ACGGAATGTA
T7 [phos]ACACCATGTA
T8 [phos]ACCAGATGTA
T9 [phos]ACCTACTGTA

T10 [phos]ACTACCTGTA
T11 [phos]ACAGTCTGTA
T12 [phos]ACGTCTTGTA
T13 [phos]ACCCCTTGTA
T14 [phos]ACTGGTTGTA
T15 [phos]ACTCAGTGTA
T16 [phos]ACGATGTGTA
T17 [phos]ACACTTGCTA
T18 [phos]ACTATCCCTA
T19 [phos]ACTCATCCTA
T20 [phos]ACGTTTCCTA
T21 [phos]ACTACTGCTA
T22 [phos]ACCAAGTCTA
T23 [phos]ACTTCGTCTA
T24 [phos]ACGAAAGCTA
T25 [phos]ACATCAGCTA
B1 GTGTACTCTACGTAGC
B2 GTGTACTCAGTGTAGC
B3 GTGTACTCCTAGTAGC
B4 GTGTACGTGAAGTAGC
B5 GTGTACGGTTTGTAGC
B6 GTGTACATTCCGTAGC
B7 GTGTACATGGTGTAGC
B8 GTGTACATCTGGTAGC
B9 GTGTACAGTAGGTAGC
B10 GTGTACAGGTAGTAGC
B11 GTGTACAGACTGTAGC
B12 GTGTACAAGACGTAGC
B13 GTGTACAAAGGGTAGC
B14 GTGTACAACCAAGTAGC
B15 GTGTACACTGAGTAGC
B16 GTGTACACATCGTAGC
B17 GTGTAGCAAGTGTAGC
B18 GTGTAGGGATAGTAGC
B19 GTGTAGGATGAGTAGC
B20 GTGTAGGAAACGTAGC
B21 GTGTAGCAGTAGTAGC
B22 GTGTAGACTTGGTAGC
B23 GTGTAGACGAAGTAGC
B24 GTGTAGCTTTCGTAGC
B25 GTGTAGCTGATGTAGC

Table 2

Oligonucleotide Combinations

A. Scheme for oligonucleotide pools. Using non-overlapping sets of oligonucleotides for the less diverse pools permits unambiguous identification of these standards with the sequence of an individual clone.

B. Scheme for combination of pools. The total sequence diversity of each combination is the product of the number of words in the left, middle and right positions. The top part of the table shows that each resulting standard can be identified by a unique restriction digestion pattern. The bottom part of the table shows that each resulting standard can be identified after sequencing of a small number of clones.

A. Pools:					
L mixtures					
Pool	Number of sequences (words)	Component L oligos	AluI	HinDIII	Amount of each component oligo (from 100 μ M stock)
La	1	40	–	+	30 μ L
Lb	8	1,2,3,4,5,6,7,8	+	–	5 μ L
Lc	14	26,27,28,29,30,31,32,33,34,35,36,37,38,39	–	–	5 μ L
Ld	25	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25	+	–	5 μ L
Le	40	All 1–40	+/-		5 μ L

T/B mixtures			
Pool	Number of sequences (words)	Component T and B oligos	Amount of each component oligo (from annealing reaction)
Ma	1	17	30 μ L
Mb	2	15,16	25 μ L
Mc	8	18,19,20,21,22,23,24,25	5 μ L
Md	14	1,2,3,4,5,6,7,8,9,10,11,12,13,14	5 μ L
Me	25	All 1–25	5 μ L

R mixtures			
Pool	Number of sequences (words)	Component R oligos	Amount of each component oligo (from 100 μ M stock)
Ra	1	17	30 μ L
Rb	2	15,16	25 μ L
Rc	8	18,19,20,21,22,23,24,25	5 μ L
Rd	14	1,2,3,4,5,6,7,8,9,10,11,12,13,14	5 μ L
Re	25	All 1–25	5 μ L

B. Combinations												
Restriction pattern												
Standard	Words	Left	Mid	Right	Left	Mid	Right	AluI	HindIII	RsaI	Sau3a	
A	1	1	1	1	La	Ma	Ra	+	+		+	
B	4	1	2	2	La	Mb	Rb	+	+	+	+	
C	16	1	2	8	La	Mb	Rc	+	+	+		
D	64	1	8	8	La	Mc	Rc	+	+			
E	128	8	8	2	Lb	Mc	Rb	+	+		+	
F	512	8	8	8	Lb	Mc	Rc	+	+			
G	896	8	14	8	Lb	Mc	Rc	+	+	+		
H	1568	14	8	14	Lc	Mc	Rd				+	
I	2744	14	14	14	Lc	Mc	Rd			+	+	
J	4900	25	14	14	Ld	Mc	Rd	+	+	+	+	
K	8750	25	25	14	Ld	Me	Rd	+	+	partial	+	
L	15625	25	25	25	Ld	Me	Re	+	+	partial	partial	
M	25000	40	25	25	Le	Me	Re	partial	partial	partial	partial	

Sequence

Sequences required to ID												
Sequences required to ID												
Standard	Words	Left	Mid	Right	Left	Mid	Right	Single	Multiple			
A	1	1	1	1	La	Ma	Ra	x				
B	4	1	2	2	La	Mb	Rb	x				
C	16	1	2	8	La	Mb	Rc	x				
D	64	1	8	8	La	Mc	Rc	x				
E	128	8	8	2	Lb	Mc	Rb		x			
F	512	8	8	8	Lb	Mc	Rc		x			
G	896	8	14	8	Lb	Mc	Rc		x			
H	1568	14	8	14	Lc	Mc	Rd	x				
I	2744	14	14	14	Lc	Mc	Rd	x				
J	4900	25	14	14	Ld	Mc	Rd		x			
K	8750	25	25	14	Ld	Me	Rd		x			
L	15625	25	25	25	Ld	Me	Re		x			
M	25000	40	25	25	Le	Me	Re		x			

Table 3
Illumina sequencing adapter oligonucleotides

The left-sided Gex adapter pairs (LT and LB) contained 13 different bar codes (shown in red). These allowed the entire set of diversity standards to be sequenced in one Illumina sequencing lane. One set of right-sided Gex adapters was ligated to all of the diversity standards. The adapters were designed to use the PCR and sequencing primers of the Illumina NlaIII Gex sequencing protocol. Oligonucleotide sequences © 2006 Illumina, Inc. All rights reserved.

NlaIII Gex Adapter 1 (left)			
Complementary oligo pairs with 5 nucleotide barcode:			
LT1	ACAGG TTCAGAGT TCTACAGTCCGACATGTGGTT	LB1	GTATAACCACATGTCGGACTGTAGAACTCTGAAC
LT2	ACAGG TTCAGAGT TCTACAGTCCGACATGGACAA	LB2	GTATTTGTCCATGTCGGACTGTAGAACTCTGAAC
LT3	ACAGG TTCAGAGT TCTACAGTCCGACATGAGGAA	LB3	GTATTTCTCATGTCGGACTGTAGAACTCTGAAC
LT4	ACAGG TTCAGAGT TCTACAGTCCGACATGGTACA	LB4	GTATTGTACCATGTCGGACTGTAGAACTCTGAAC
LT5	ACAGG TTCAGAGT TCTACAGTCCGACATGCATCA	LB5	GTATTGATGCATGTCGGACTGTAGAACTCTGAAC
LT6	ACAGG TTCAGAGT TCTACAGTCCGACATGACCTA	LB6	GTATTAGGTCATGTCGGACTGTAGAACTCTGAAC
LT7	ACAGG TTCAGAGT TCTACAGTCCGACATGGGTTA	LB7	GTATTAACCCATGTCGGACTGTAGAACTCTGAAC
LT8	ACAGG TTCAGAGT TCTACAGTCCGACATGCTGTA	LB8	GTATTACAGCATGTCGGACTGTAGAACTCTGAAC
LT9	ACAGG TTCAGAGT TCTACAGTCCGACATGTGAGA	LB9	GTATTCTCACATGTCGGACTGTAGAACTCTGAAC
LT10	ACAGG TTCAGAGT TCTACAGTCCGACATGATCAC	LB10	GTATGTGATCATGTCGGACTGTAGAACTCTGAAC
LT11	ACAGG TTCAGAGT TCTACAGTCCGACATGTCTAC	LB11	GTATGTAGACATGTCGGACTGTAGAACTCTGAAC
LT12	ACAGG TTCAGAGT TCTACAGTCCGACATGCTAGT	LB12	GTATACTAGCATGTCGGACTGTAGAACTCTGAAC
LT13	ACAGG TTCAGAGT TCTACAGTCCGACATGTACTC	LB13	GTATGAGTACATGTCGGACTGTAGAACTCTGAAC
NlaIII Gex Adapter 2 (right)			
5' ATCATCGTATGCCGTCTTCTGCTTG			
5' CAAGCAGAAGACGGCATAACGA			
Gex PCR Primer 1 (3' phosphorothioate protected)			
5' CAAGCAGAAGACGGCATAACGA			
Gex PCR Primer 2 (3' phosphorothioate protected)			
5' AATGATACGGCGACACCGACAGGTTTCAGAGTCTACAGTCCGA			
Gex Sequencing Primer			
5' CCGACAGGTTTCAGAGTCTACAGTCCGACATG			