

A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing

Poornima Parameswaran¹, Roxana Jalili³, Li Tao⁴, Shadi Shokralla³,
Baback Gharizadeh³, Mostafa Ronaghi³ and Andrew Z. Fire^{2,*}

¹Department of Microbiology and Immunology, ²Department of Pathology and Department of Genetics, Stanford University School of Medicine, ³Stanford Genome Technology Center and ⁴Department of Biological Sciences, Stanford University, Stanford, CA-94305, USA

Received July 30, 2007; Revised September 10, 2007; Accepted September 11, 2007

ABSTRACT

Multiplexed high-throughput pyrosequencing is currently limited in complexity (number of samples sequenced in parallel), and in capacity (number of sequences obtained per sample). Physical-space segregation of the sequencing platform into a fixed number of channels allows limited multiplexing, but obscures available sequencing space. To overcome these limitations, we have devised a novel barcoding approach to allow for pooling and sequencing of DNA from independent samples, and to facilitate subsequent segregation of sequencing capacity. Forty-eight forward–reverse barcode pairs are described: each forward and each reverse barcode unique with respect to at least 4 nt positions. With improved read lengths of pyrosequencers, combinations of forward and reverse barcodes may be used to sequence from as many as n^2 independent libraries for each set of ' n ' forward and ' n ' reverse barcodes, for each defined set of cloning-linkers. In two pilot series of barcoded sequencing using the GS20 Sequencer (454/Roche), we found that over 99.8% of obtained sequences could be assigned to 25 independent, uniquely barcoded libraries based on the presence of either a perfect forward or a perfect reverse barcode. The false-discovery rate, as measured by the percentage of sequences with unexpected perfect pairings of unmatched forward and reverse barcodes, was estimated to be <0.005%.

INTRODUCTION

With the advent of high-throughput sequencing, characterization of the nucleic acid world is proceeding at an accelerated pace. Three major high-throughput sequencing platforms are in use today: the Genome Sequencers from Roche/454 Life Sciences [GS-20 or GS-FLX; (1)], the 1G Analyzer from Illumina/Solexa (2) and the SOLiD System from Applied Biosystems (<http://solid.appliedbiosystems.com>). Comparison across the three platforms reveals a trade-off between average sequence read length and the number of DNA molecules that are sequenced. At present, the Solexa and SOLiD systems provide many more sequence reads, but render much shorter read lengths than the 454/Roche Genome Sequencers. This makes the 454/Roche platform appealing for use with barcoding technology, as the enhanced read length facilitates the unambiguous identification of both complex barcodes and sequences of interest.

The 454/Roche Genome Sequencers are called pyrosequencers because their sequencing technology is based on the detection of pyrophosphates released during DNA synthesis (3). A few sequencing runs using 454/Roche's pyrosequencing platform can generate sufficient coverage for assembling entire microbial genomes (4,5), for the discovery, identification and quantitation of small RNAs (6,7), and for the detection of rare variations in cancers (8), among many other applications. For analysis of multiple libraries, the currently available 454/Roche pyrosequencers can accommodate a maximum of only 16 independent samples, which have to be physically separated using manifolds on the sequencing medium. These separation manifolds occlude wells on the

*To whom correspondence should be addressed. Tel: +1 650 723 2885; Fax: +1 650 724 9070; Email: afire@stanford.edu

© 2007 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

sequencing plate from accommodating bead-bound DNA template molecules, and thus restrict the number of output sequences. To overcome these limitations, we describe a high-information-content barcoding approach in which each sample is associated with two uniquely designed, 10-nucleotide barcodes. The presence of these assigned barcodes allow for independent samples to be pooled together for sequencing, with subsequent bioinformatic segregation of the pyrosequencer output. By not relying on physical separators, this procedure maximizes sequence space and multiplexing capabilities.

'Barcodes', or unique DNA sequence identifiers, have historically been used in several experimental contexts. In sequence-tagged mutagenesis (STM) screens, a sequence barcode acts as an identifier or type specifier in a heterogenous cell-pool or organism-pool. STM barcodes are usually 20–60 nt long, are pre-selected or follow ambiguity codes, and are present as one unit or split into pairs (9). Such long barcodes are not ideally suited for use with available pyrosequencing platforms because of restrictions on read length. Several recent papers have reported the use of very short (2- or 4-nt) barcodes with pyrosequencing platforms (6,10,11). Although such approaches are valuable, there are applications that call for a more definitive assignment of samples and/or for enhanced multiplexing capabilities. We describe here the design and testing of a series of 10-base barcodes that differ from each other at multiple positions (by at least 4 nt), and have a common specification that facilitates resolution of ambiguities. Using the GS20 platform (454/Roche), these barcodes allowed us to simultaneously pyrosequence small RNA libraries from 25 diverse samples for each pilot run, and unambiguously assign 99.8% of obtained barcoded sequences by bioinformatically probing for an error-free barcode at either end of the sequence of interest. The false-discovery rate, defined as rate of occurrence of mispaired forward and reverse barcodes, was 0.00042% (~1 in 237 000 sequences) in run I, and 0.0044% (~12 in 271 000 sequences) in run II. Discussed here are the modifications made for cloning and amplification protocols to handle highly complex nucleic acid pools, the considerations used in designing the barcodes, and the efficacy of the barcoding technology. We also speculate on the use of these pyrosequencer-tailored barcodes with other sequencing platforms.

MATERIALS AND METHODS

Cloning of small RNAs

We sought to identify novel small RNAs in 50 different RNA preparations derived from a diverse set of virus-infected and control cell and tissue samples. To pursue this analysis, small RNA fractions from each cell or tissue sample were independently isolated (using the miRvana small RNA isolation protocol (Ambion)), and processed to produce 50 individually barcoded cDNA libraries that could be sequenced using the Roche/454 platform. cDNA libraries were constructed (using slight modifications of standard procedures) as follows: RNA 3' termini were coupled to adenylated DNA linkers as described (12,13),

and subsequent ligation of the second linker was performed either directly to the 5' termini of the RNAs [5'-P-dependent cloning; (12)], or to the 3' termini of the cDNAs synthesized from the 3'-linkered RNAs [5'-P-independent cloning; (13)]. The following sets of linkers were used for cloning:

linkerA (RNA–DNA hybrid linker): ATCGTrArGrGrCrArCrCrUGrArArA

linkerB (IDT Linker 2): rAppCACTCGGGCACCAAGGA/3ddC/

linkerC (DNA linker): pCACTCGGGTGCCAAGGA/3ddC/

linkerD (IDT Linker 1): rAppCTGTAGGCACCATCAAT/3ddC/

Ribonucleotides are marked by the presence of 'r' before the nucleotide. 'App' is adenylated nucleotide; 'ddC' is dideoxycytidine and 'p' is phosphate group. Both linkerB and linkerD may be used with either cloning protocol. For our experiments, linkerA and linkerB were used for cloning 5'-monophosphated small RNAs, while linkerC and linkerD were used in constructing libraries using the 5'-P-independent cloning protocol. Published protocols for library construction (12,13) were modified to reduce the number of gel purifications. In particular, size fractionation on a 12% denaturing PAGE–urea gel was exclusively performed after the first ligation event. This substantially reduced the required amount of starting material (from several microgram to as little as 0.25 µg of the small RNA fraction), and minimized loss of material that would have otherwise occurred with the two additional gel elutions.

Overall design of barcoded primers used in amplifying cDNA libraries

The barcoded primers were each 45–46 nt long and consisted of: (i) the 454 amplicon adapter (the amplicon sequencing primer annealing site; the 'F-adapter or R-adapter'), (ii) the barcode (F-barcode or R-barcode) and (iii) the F-Cloning-Linker or R-Cloning-Linker (the sequences corresponding to the linkers used for construction of the initial nucleic acid library) (Figure 1).

Blueprint for barcodes

FORWARD barcode format: HIJJK LLMNO

REVERSE barcode format: HIJJK LMMNO

The quartets (H, I, J and K) and (L, M, N and O) each independently map to the four nucleotides (A, C, G or T). Guidelines for positional nucleotide restrictions in the barcodes are presented in Table 1.

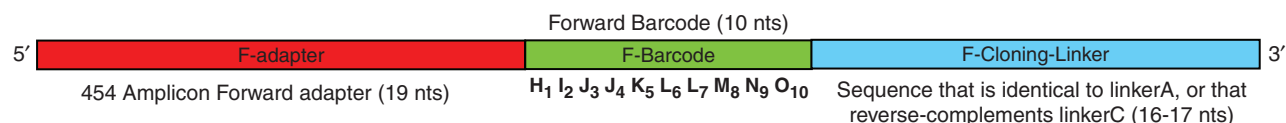
Primer sequence corresponding to adjusted linker sequence (linker-primer)

linker-primerA (corresponds to linkerA): ATCGTAGG CACCTGAAA

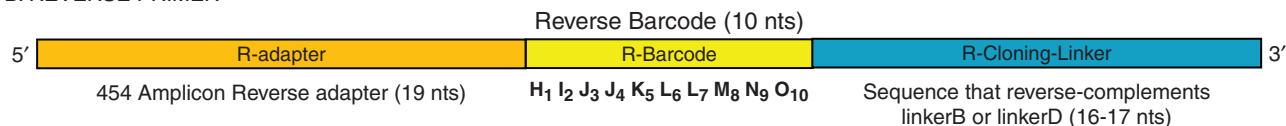
linker-primerB (corresponds to linkerB): ACCGTCCTTG GTGCCCCG

linker-primerC (corresponds to linkerC): TCCTTGGCAC CCGAGTG

A. FORWARD PRIMER



B. REVERSE PRIMER



C. SEQUENCING READS: FORWARD & REVERSE

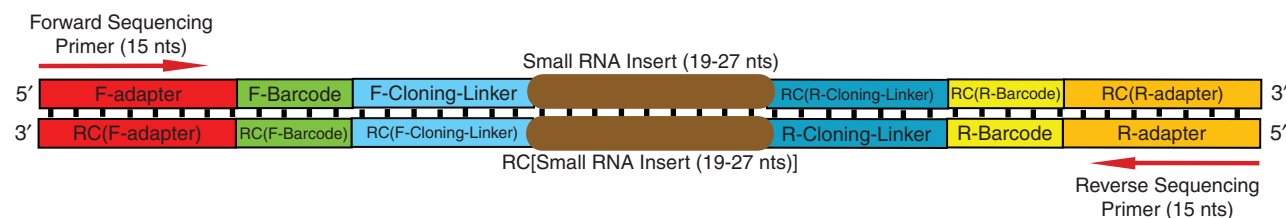


Figure 1. Design of forward and reverse primers. The synthesized primers are 45–46 nt long. (A and B) The template for the primers is: 454-Adapter::Barcode::Linker-Primer. Individual specifications for forward and reverse barcodes are also indicated. (C) Diagrammatic representation of the GS20 forward and reverse sequencing reads. ‘RC’ stands for Reverse Complement. Teeth denote base pairing.

Table 1. Guidelines for barcodes

Restrictions	Cloning protocol	Guidelines for restrictions
$H \neq (I, J, K)$ $I \neq (J, K)$ $J \neq K$ $K \neq L$ $L \neq (M, N, O)$ $M \neq (N, O)$ $N \neq O$	5-P and DL	Adjoining nucleotides must be different (except at dinucleotide positions). For each barcode, (H, I, J, K) and (L, M, N, O) uniquely map to one of (A, C, G, T). Each barcode has two distinct dinucleotide pairs: at positions (3,4) and (6,7) for the forward barcodes, and at positions (4,5) and (7,8) for reverse barcodes.
$H \neq G$	5-P and DL	The terminal nucleotide of the F-adapter and R-adapter (Figure 1) is a guanosine (in both (5'-phosphate-dependent and 5'-phosphate-independent protocols); ‘H’ cannot be part of a dinucleotide, and hence cannot be a guanosine.
$O \neq A$	5-P	First nucleotide of F-cloning-linker and R-cloning-linker is an adenosine (in the 5'-phosphate-dependent protocol). ‘O’ cannot be part of a dinucleotide, and hence cannot be an adenosine.
$O \neq T$	DL	First nucleotide of F-cloning-linker and R-cloning-linker is a thymidine (in the 5'-phosphate-independent protocol). ‘O’ cannot be part of a dinucleotide, and hence cannot be a thymidine.

Barcodes used with forward and reverse primers have different designs, with specific restrictions imposed at every nucleotide position. These restrictions are listed, and further discussed in the text. Numbers indicate positions of nucleotides in the barcode. 5-P: 5'-phosphate-dependent cloning; DL: 5'-phosphate-independent cloning.

FORWARD barcode format: $H_1 I_2 J_3 J_4 K_5 L_6 L_7 M_8 N_9 O_{10}$.

REVERSE barcode format: $H_1 I_2 J_3 K_4 K_5 L_6 M_7 M_8 N_9 O_{10}$.

linker-primerD (corresponds to linkerD): TTGATGGT GCCTACAG

DNA amplification and pooling of independent samples

The linkered cDNA pool from each individual sample was amplified for 20 cycles [as described in (12)] with an unique pair of barcoded primers. A second round of amplification with temperatures and cycle lengths identical to the initial round, but with variable number

of total cycles (six, eight and ten), was performed using 1/25th of the first PCR product, dNTPs, barcoded primers, buffer and *Taq* DNA polymerase in a 40 µl PCR mix. An analytical, non-denaturing 4% Nusieve gel was used to visually determine the fewest number of cycles that yielded a visible signal for PCR products containing inserts corresponding to small RNAs. DNA was extracted from the gel using the Qiaquick gel extraction kit (Qiagen). Although the instructions for gel extraction called for dissolution of the gel slice in

Buffer QG at 50°C, we performed all manipulations of the DNA-containing buffers at room temperature, to ensure that the complex PCR mixtures were not denatured.

To obtain a rough quantitation of individual libraries, 10% of each recovered PCR product was separated on an analytical 2% agarose gel. Based on the intensities of bands on this gel, PCR products from 25 samples were pooled together in roughly equimolar ratios. Quality of the mixture was tested by conventional Sanger sequencing of 46 PCR products that had been cloned into TOPO-TA vectors (Invitrogen). If the sequence quality was satisfactory, 300 ng of the mixture was run on a non-denaturing 2% low melting point (LMP)-agarose gel (run I) or on a denaturing 6% PAGE-urea gel (run II; recommended) for quantification, and for further size-restricted purification of barcoded species that contain small RNA inserts. DNA was eluted from the PAGE-urea gel overnight at 4°C in 0.3 M NaCl.

Processing of pooled sample for high-throughput sequencing

Amplicon libraries prepared as described above may be entered directly into the 454/Roche sequencing pipeline, with the only major variable being the amount of DNA from the pooled sample used to seed the emulsion PCR reaction. Approximately 50 ng of the purified pool was set aside for each pyrosequencing run. For run I, DNA was not heat-denatured, and concentrations of both single-stranded (using Agilent 2100 Bioanalyzer's RNA chip assay; Agilent Technologies) and double-stranded (using the Quant-iT PicoGreen dsDNA Assay; Invitrogen) species were measured, before the DNA was channeled into the Roche/454 Amplicon sequencing pipeline. As estimating concentrations of appropriately sized molecules in each library has been most reliable following denaturation of samples, for run II (and for subsequent uses of this technology), we denatured input DNA (by flash heating and cooling) prior to both measurement of concentration (using the 2100 BioAnalyzer) and emulsion PCR. Post-denaturation, different ratios of DNA molecules to DNA-capturing beads were tested in an initial attempt to find a concentration of DNA that would result in a 10–30% bead recovery after emulsion PCR (1). For run II of this analysis, the DNA to bead ratio used was 3:1. All subsequent procedures were carried out essentially as described in the GS20 guide-book for amplicon sequencing (454/Roche).

Data handling

Program 'Barsort' was used to resolve sequences into separate data bins based on perfect match to the 10-nt barcode sequence + adjoining 'x' nucleotides of the linker sequence ($0 \leq x \leq \text{length of linker}$). Insert sequences were then mapped to RNA databases using standard pattern-matching algorithms. The false-discovery rate was estimated based on sequence segregation into various subcategories using the program 'Barverify'. Barsort and Barverify are available as source code and executables on request.

RESULTS AND DISCUSSION

Design of barcoded primers

We have devised a barcoding approach to substantially enhance the scope and capacity of multiplexed high-throughput pyrosequencing. Specifications for the barcodes are based on details of pyrosequencing error characteristics. Sets of barcodes have been independently designed for the forward PCR primers (FORWARD barcodes) and the reverse PCR primers (REVERSE barcodes) used in library construction (Figure 1). Length of the barcode was chosen to allow for definitive assignment of barcodes, while minimizing the number of base reads used per barcode. The barcode design (Table 1) facilitated unambiguous assignment, with care taken to avoid homopolymeric runs that are particularly problematic in pyrosequencing technologies. All forward barcodes were of the type HIJJK LLMNO, while reverse barcodes were of the type HIJJK LMMNO. H, I, J, K and L, M, N, O are uniquely partnered with one of four nucleotides: A, C, G and T. The barcodes thus contained two instances of a 1–2 base run of each nucleotide (one from the H, I, J, K set, and the other from the L, M, N, O set), with no more than two successive occurrences of the same nucleotide (Table 1). This design avoids significant problems due to error rate at homopolymeric runs of nucleotides. The forward and reverse barcodes varied in dinucleotide positioning along the length of the barcode. Positions 3, 4 (nucleotide J) and 6, 7 (nucleotide L) were homopolymeric dinucleotides in the forward barcode, while positions 4, 5 (nucleotide K) and 7, 8 (nucleotide M) were homopolymeric dinucleotides in the reverse barcode.

There were several restrictions at specific positions in the barcode. The start position of the barcode (position H) was required to differ from the last nucleotide of the preceding adapter to avoid homodinucleotides at the Adapter::Barcode junctions (Figure 1). The terminal nucleotide of the barcode (position 'O') was constrained to prevent identity with the first nucleotide of both the F-Cloning-Linker and the R-Cloning-Linker. To ensure identical nucleotide possibilities at 'O' for all forward and reverse barcodes, priming sequences that correspond to the F-Cloning-Linker and to the R-Cloning-Linker were adjusted as necessary so that they started with the same base (refer to Methods section). These adjustments avoided any unnecessary restrictions on the last base of the barcode ('O'), created independent sets of restrictions for positions H and O, eliminated the possibility for a dinucleotide at the Barcode::Cloning-Linker junctions and at the Adapter::Barcode junctions, and increased the total number of forward–reverse barcode pairs.

Sixty-four barcode pairs were thus designed, with any two forward or reverse barcodes differing in at least 4 nt: one doublet + two singlets (Table 2). These serve as strong unique identifiers for each of the forward and reverse barcode sets. Due to the terminal nucleotide restriction (on position 'O'), only 48 barcodes may be used with a single 5'–3' pair of linkers used in cloning. We make one additional note on the choice of barcode sequences: in applications where there is a need for fully optimized

Table 2. List of possible barcodes

Forward barcodes	Reverse barcodes
AGCCTAAGCT	AGCTTAGGCT
AGTTCAAGTC	AGTCCAGGTC
ACTTGAAGTG	ACTGGACCTG
ACGGTAACGT	ACGTTACCGT
ATCCGAATCG	ATCGGATTCG
ATGGCAATGC	ATGCCATTGC
CAGGTCCAGT	CAGTTCAAGT
CATTGCCATG	CATGGCAATG
CTAAGCCTAG	CTAGGCTTAG
CGAATCCGAT	CGATTCCGAT
TCAAGTTAGC	TAGCCTAAGC
TACCGTTACG	TACGGTAACG
TGAACTTGAC	TGACCTGGAC
TAGGCTTCAG	TCAGGTCCAG
AGCCTCCAGT	AGCTTCTTAG
AGCCTGGCAT	AGCTTGCCAT
AGTTCCGACT	AGTCCGAAGT
AGTTCTTGAC	AGTCCTCCAG
ACGGTCCATG	ACGTTCAATG
ACTTGTTTAC	ACTGGTGGAC
ACTTGCCGAT	ACTGGCGGAT
ACGGTGGATC	ACGTTGAATC
ATCCGCCTAG	ATCGGCAAGT
ATCCGTTAGC	ATCGGTAAGC
ATGGCGGTAC	ATGCCGTTAC
ATGGCTTACG	ATGCCTAACG
CAGGTAAGTC	CAGTTAGGTC
CAGGTGGCAT	CAGTTGCCAT
CATTGAAGCT	CATGGAGGCT
CTAAGTTCAG	CTAGGTCCAG
CTAAGAACGT	CTAGGACCGT
CTGGATTGAC	CTGAATGGAC
CATTGTTAGC	CATGGTAAGC
CTGGAGGACT	CTGAAGAACT
CGAATAACTG	CGATTACCTG
CGAATGGATC	CGATTGAATC
CGTTAGGTAC	CGTAAGTTAC
CGTTATTACG	CGTAATAACG
TAGGCAAGCT	TAGCCAGGCT
TACCGCCATG	TACGGCAATG
TACCGAAGTC	TACGGAGGTC
TGAACGGCAT	TGACCGCCAT
TGAACAATCG	TGACCATTCTG
TGCCACCGAT	TGCAACGGAT
TGCCAGGACT	TGCAAGAACT
TCGGACCTAG	TCGAACCTAG
TCAAGCCAGT	TCAGGCAAGT
TCAAGAATGC	TCAGGATTGC
CTGGACCTGA	CTGAACCTGA
CGTTACCGTA	CGTAACGGTA
TGCCATTGCA	TGCAATGGCA
TCGGATTCTGA	TCGAATCCGA
AGCCTGGCTA	AGCTTGCCTA
AGCCTCTCTGA	AGCTTCTTGA
ACTTGTTCTGA	ACTGGTCCGA
ATCCGCCGTA	ATCGGCGGTA
ATCCGTTCTGA	ATCGGTCCGA
ATGGCGGTCA	ATGCCGTTCA
CAGGTGGCTA	CAGTTGCCTA
CTAAGTTGCA	CTAGGTGGCA
CGTTAGGTCA	CGTAAGTTCA
TCGGACCGTA	TCGAACGGTA
TCAAGCCTGA	TCAGGCTTGA
TGCCAGGTCA	TGCAAGTTCA

The barcodes in this compilation have restrictions on the first nine bases only. Restrictions on the terminal base are applied after selection and adjustment of 5'–3' cloning linker pairs.

read length, an additional preference may be imposed for barcodes whose sequences best utilize the sequencing order (TACG for the GS20) so that a few more sequencing cycles may be available for the DNA of interest.

Testing barcoded primers

The barcoded primers were 45–46 nt in length (Supplementary Table 1), had long overhangs (29–30 nt), and short regions of complementarity to the template (16–17 nt; did not span the entire Cloning-Linker sequence). Despite the extended length of the barcoded primers, complex pools of cDNA containing 19–27 nt small RNA inserts were efficiently amplified to yield PCR products that were 115–123 nt long.

Deciphering the nature of products amplified from complex pools of cDNA

Construction and sequencing of libraries from complex mixtures of RNA templates involves numerous steps that may skew the relationship between ratios of molecules in the initial pool and obtained ratios of corresponding sequences. Differential efficiencies of RNA extraction, linker addition and reverse transcription may skew the initial cDNA pool. Differences in amplification efficiencies of individual sequences in the cDNA pool, however slight, may bias the library population over many rounds of PCR. Sequence-specific differences in efficiency or fidelity of sequencing may distort the final proportions of assigned DNA sequences. These biases have been extensively discussed in literature; although potentially confounding, their effects may be controlled for to some extent, so that sequence representations in the cDNA and subsequent PCR-amplified libraries at least provide for an initial assessment of sequence incidence in the original small RNA populations.

We note one additional concern that is highly relevant to preparation of libraries for high-throughput sequencing. This concern specifically relates to purification and handling of pooled DNA products following amplification of complex cDNA libraries. Amplified DNA pools generally consist of a mixture of single strands, perfect duplexes and partial duplexes. The overall structural composition of the amplified DNA pool is determined by the extent of amplification in the last round of PCR. DNA molecules that were substrates for polymerase extension during the last round of PCR will be double-stranded. Molecules that fail to extend in the last round of PCR may remain single-stranded, may form heteroduplexes (if they anneal to a distinct molecule during the last PCR cycle; Figure 2), or may form perfect duplexes (if they find and anneal to an exact complement during the last PCR cycle). This heterogeneity will be more prominent in situations where extension has become inefficient (e.g. PCR taken beyond a point of saturation), resulting in the majority of molecules being single stranded or heteroduplex during thermal cycling. Only abundant single-stranded species would be more likely to reanneal to form homoduplexes (Figure 1c). The low probability for rare single-stranded species to find perfect partner strands pushes them towards reannealing with

A. HETERODUPLEX CLASS A



B. HETERODUPLEX CLASS B

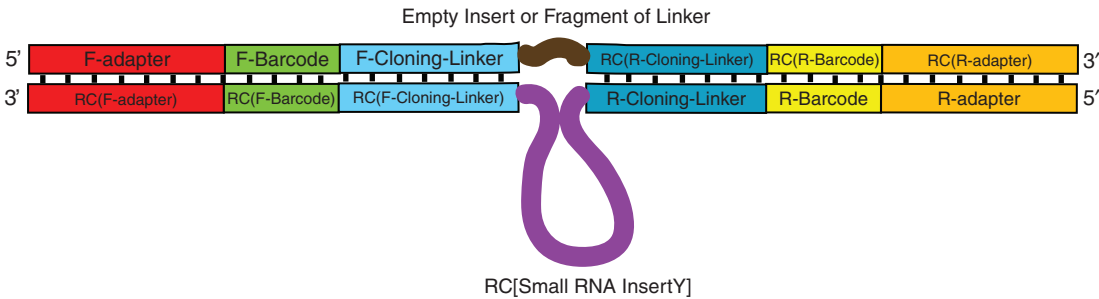


Figure 2. A diagrammatic representation of heteroduplexes that may be formed in an amplified pool. (A) Heteroduplexes formed between molecules from the same sample that have the same barcode, but different RNA inserts (X and Y). (B) Heteroduplexes formed between molecules with RNA inserts and molecules with no RNA inserts (or with fragments of linkers as inserts). Formation of these unusual duplexes may be facilitated by the 45–46 nt complementarity at either end of the insert. Thus, three types of molecules may be present during later stages of PCR: single-stranded, perfectly double stranded (Figure 1C) and heteroduplexes (shown here). The ratio of the three species is determined by the number of PCR cycles. As in Figure 1, ‘RC’ stands for Reverse Complement, and teeth denote base pairing.

non-identical PCR products (i.e. with library elements that have distinct inserts), resulting in the formation of heteroduplex structures (Figure 2).

The heterogenous mixture of single strands, perfect duplexes and heteroduplexes skews the sequence incidence only if there is a step in the procedure that enriches for double-stranded molecules post-PCR. This is the case, however, with many gel isolation procedures that are employed to size-select for PCR products in a pre-defined size range. Biased selection in these protocols is attributed to a combination of factors: (i) differences in mobility between single-stranded, double-stranded and heteroduplex molecules on the gel, (ii) differential recovery from gel-isolation procedures, which may enrich for segments that have either formed perfect duplexes, or are capable of forming perfect duplexes.

Our observations confirm that the overall nature of the DNA pool number was influenced by the extent of amplification, i.e. by the number of PCR cycles (Figure 3), since a migration shift on the agarose gel was evident upon the addition of two to four PCR cycles (with all other PCR conditions being constant). This shift is consistent with a transition between duplex structures (a relatively tight band), and a mixture of heteroduplex and single-stranded material (a more diffuse band starting above the presumed duplex band). In order to maximize the accuracy of sequence representation for each individual library, we took samples from PCR cycles prior to saturation. Products migrating with the expected double-stranded DNA population were then isolated from an agarose gel, recovered (by extraction without denaturation) and quantitated, before they were used

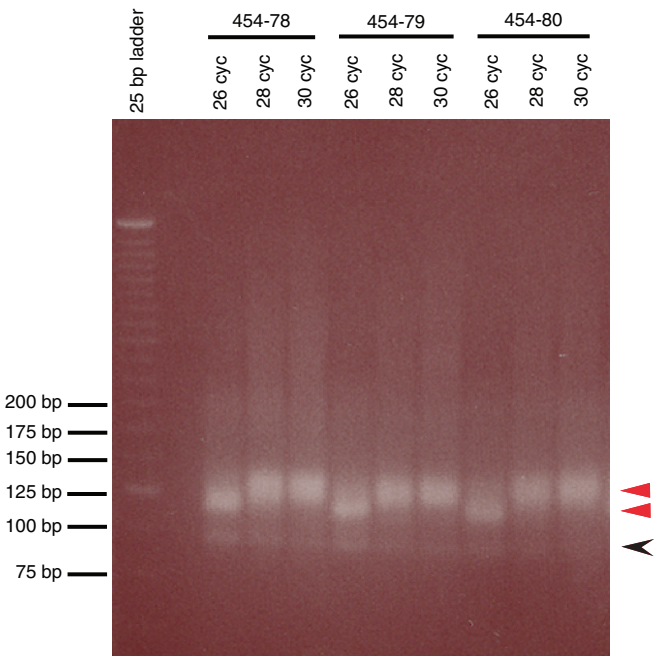


Figure 3. Visualizing the nature of amplified products from various cycles of PCR. With an increase in cycle number (twenty cycles of an initial round of PCR followed by six, eight or ten cycles of a second round of PCR), there is an evident shift in mobility of PCR products that contain a small RNA insert. Effect of PCR cycle number on three different samples is shown, stressing the importance of titrating the total number of DNA amplification cycles, to avoid saturation of the PCR amplification. Red arrows represent the two sizes of the insert-containing PCR products. PCR products without small RNA inserts migrate as faint bands between 75 and 100 bp (black arrow).

Table 3. Summary of sequence distribution for both pyrosequencing runs as a function of barcode motif length

Run number	Search motif	Perfect 5' and 3' Motifs	Mismatched 5' and 3' motifs	Perfect 3' and imperfect 5' motifs	Perfect 5' and imperfect 3' motifs	Duplicate perfect 5' motifs	Duplicate perfect 3' motifs	Total number of sequences
I	Barcode (10 nt)	229 277	2349	2323	2923	1	0	23 6873
I	Barcode (10 nt) + Cloning-Linker (1 nt)	226 785	295	4339	5211	2	233	23 6865
I	Barcode (10 nt) + Cloning-Linker (2 nt)	226 108	4	5368	5858	0	0	23 7338
I	Barcode (10 nt) + Cloning-Linker (3 nt)	224 645	3	5800	6892	0	0	23 7340
I	Barcode (10 nt) + Cloning-Linker (4 nt)	223 744	3	6258	7335	0	0	23 7340
I	Barcode (10 nt) + Cloning-Linker (COMPLETE)	205 020	1	12 095	20 224	0	0	23 7340
II	Barcode (10 nt)	258 929	5061	3138	3736	2	179	27 1045
II	Barcode (10 nt) + Cloning-Linker (1 nt)	257 998	17	6128	7160	3	1	27 1307
II	Barcode (10 nt) + Cloning-Linker (2 nt)	257 150	12	6587	7557	2	1	27 1309
II	Barcode (10 nt) + Cloning-Linker (3 nt)	255 695	11	7175	8427	0	1	27 1309
II	Barcode (10 nt) + Cloning-Linker (4 nt)	254 491	11	7818	8988	0	1	27 1309
II	Barcode (10 nt) + Cloning-Linker (COMPLETE)	226 041	11	14 662	30 595	0	0	27 1309

This table summarizes the data in Supplementary Tables 4 and 5. Fewer sequences with mismatched 5' and 3' motifs (i.e. sequences that do not have corresponding 5' and 3' motifs) were obtained if the lengths of the search motif used as unique identifiers for the sequences were increased. The number of sequences with imperfect motifs also correspondingly increased with the increased length of the barcode motif, reflective of the high error rate of the pyrosequencing technology. There were also rare instances of sequences with two different 3' motifs or two different 5' motifs. These may be PCR artifacts. *Terminology:* 5' barcodes may be of the forward or reverse category (depending on the sequencing primer used), and are in the 5' flank of the read. 3' barcodes are reverse complements of forward and reverse barcodes (depending on the sequencing primer used), and are in the 3' flank of the read.

for sequencing. Our samples were found to consist of approximately equal numbers of double-stranded and single-stranded DNA molecules by the Quant-iT PicoGreen dsDNA assay (Invitrogen), and the RNA BioChip assay (Agilent 2100 Bioanalyzer, Agilent Technologies), respectively (data not shown). Our efforts to avoid saturation of the PCR amplification, although not perfect, helped to maximize the proportion of homoduplexes formed in the last cycle of PCR, thus facilitating the recovery of both rare and common species in the library pool. All these observations stress the importance of close monitoring of complex DNA pools to ensure that overall quality is not compromised through excessive amplification of seed populations, and through subsequent manipulations of amplified populations.

Analysis of data from two independent pyrosequencing runs

DNA from two independent sets of 25 barcoded libraries (prepared with 19–27 nts RNAs from different *in vitro* and *in vivo* virus-infected systems) were pooled for sequencing runs, and the data was used to evaluate the efficacy of the barcoding technology. Both 237 639 and 271 777 sequences from runs I and II, respectively, passed the GS20's quality control filters. Since six of the forward barcodes are exact reverse complements of six reverse barcodes, we added a number of adjoining nucleotides in the Cloning-Linker (0–17) to the barcode query to reduce ambiguity while segregating the sequences into independent datasets. These barcode + linker sequences are referred to herein as 'search motifs' or 'barcode motifs'. Using 13-nt search motifs, Barsort and Barverify identified that 94.53% (run I) and 94.08% (run II) of the sequences had perfect forward and reverse barcode motifs, 2.90% (run I) and 3.24% (run II) of the sequences had perfect barcode motifs only at the 5' end of the read, and 2.44% (run I) and 2.51% (run II) had perfect barcode

motifs only at the 3' end of the read (for data and terminology, refer Table 3).

Sequences that have a perfect match to only one of the expected motifs have been further subdivided into three classes based on whether the motif at the other end has: (i) perfect identity to an unexpected motif, resulting in an unassigned 5'–3' motif combination (perfect, unexpected occurrence), (ii) a high-scoring partial match to the expected motif for that library (imperfect, expected occurrence) and (iii) no strong match to any barcode motif (imperfect, ambiguous occurrence). Perfect, 'unexpected' occurrences may be attributed to motifs with rare coincidental sequence matches (due to multiple fortuitous errors during barcode synthesis or sequencing), to contamination in PCR primer stocks, or to cross-contamination of template DNAs in the second round of PCR. The imperfect matches may arise due to errors in sequencing [particularly at the starts and ends of the sequencing reads; (1)], due to errors during conventional and/or emulsion PCR, or due to imperfections during the chemical synthesis of barcoded primers. With the barcode design used in this work (i.e. no homopolymer beyond 2 nt), we expected to minimize errors related to pyrosequencing. Indeed, the experimentally observed spectrum of errors did not match the types of homopolymer-related errors most frequently observed for pyrosequencing. Instead, the observed error spectrum appeared most closely to match that expected as a result of known imprecisions in the chemical synthesis of DNA primers used in standard PCR reactions (Supplementary Table 3). Adding to any potential protocol-related possibility of sample misassignment, we note that the highly sensitive nature of high-throughput sequencing needs to be considered in all analyses. In particular, any possible source for sample contamination upstream (or during) library construction needs to be carefully monitored.

For bioinformatic segregation of a barcoded dataset, a decision needs to be made as to how much of the

(fixed) linker sequence will be required in addition to the (variable) barcode for definitive sample assignment. The false-discovery rate, defined by the percentage of 'perfect and unexpected' occurrences, was observed to decrease to <0.005% (Table 3), if three bases of the linker were used in addition to the 10-base barcode to assign sample identities. Thus a search motif of 13 nt was determined to be ideal for our studies in minimizing false-discovery rate, while maximizing the number of sequences with error-free hits identified by the program. Sequences segregated using motifs of this length were used for all further analyses. The low false-discovery rate for perfect matches strengthens the argument that any sequence with both 5' and 3' barcode motifs will be placed with reasonable accuracy into the correct bin. For our purposes, all sequences with a match to one of the two expected sequence motifs were considered as true members of the corresponding dataset. Run I (0.13%) and run II (0.17%) of sequences had no perfect matches to any of the sequence motifs (forward or reverse), and were discarded. Thus, data from two independent sequencing runs on pools of barcoded material from a total of 50 sample libraries (25 sample libraries per run) show similar patterns with respect to sequence recovery, binning and rate of errors that result in false discovery and misassignment.

Distribution of sequences across samples

The barcoding approach requires being able to establish a relationship between ratios of material from various libraries, and the ratios of obtained sequences that correspond to those libraries. Our pooling of equimolar ratios of DNA from independent libraries based on visual quantification was evidently less than perfect, as there was variance in the representation of individual libraries in the overall pool (Supplementary Tables 4D and 5D). If the pooling was indeed equimolar, the expected mean number of sequences per independent pool would be 9596 (run I) and 10 852 (run II), with expected binomial SD of 95.98 (run I) and 102.07 (run II). Instead, in run I, we observed a SD of 2909.93 from the mean, and the number of sequences per set varied from 2950 to 17 922. In run II, we observed a SD of 2801.64 from the mean, and the number of sequences per set varied from 5347 to 15 901. The observed fluctuations between samples are not statistical in nature and may primarily be due to inconsistencies in ethidium bromide (EtBr)-based quantitation of DNA. Much of the difference may result from subjectivity in visual quantification (2-fold differences or less are considered appropriate for this method). In addition, EtBr will fluoresce more effectively with double-stranded species than with heteroduplex or single-stranded species. Thus, ratio of single-stranded to duplex species may influence the intensity of the band during gel electrophoresis, and may contribute to the observed variation. Finally, the fraction of heteroduplexes and non-annealed single strands present in the pool has an impact on the compactness of the band, and this in turn may affect precise concentration approximation. Denaturing materials from individual samples, and measuring

their concentrations using Agilent's 2100 BioAnalyzer (before pooling samples together) may overcome these drawbacks of visually estimating DNA concentrations.

CONCLUSION

We have described a pyrosequencing-tailored barcoding approach that allows for the unambiguous assignment of nucleic acid sequences from a mixture of libraries from up to 48 different samples. Lengthening the barcodes, or variations in the fixed forward and reverse linkers used to generate the initial cDNA libraries can easily expand the multiplexing capacity. A remarkably low misassignment rate (0.00042% and 0.0044% for two independent experiments) was obtained by requiring a match to the 10-base barcodes and additional three bases of the fixed linker. The low misassignment rate also indicates the potential utility of combinations of forward and reverse barcodes to generate up to 48^2 (= 2304) combination barcodes for this set of sequences.

Previously, multiplexed pyrosequencing using the 454 sequencing technology was implemented to analyze sequences from several datasets simultaneously (6,10,11). However, the barcodes used in these experiments are too short to allow for massively parallel runs involving large numbers of sample libraries. Shorter barcodes also have a steeper trade-off between number of possible barcodes and the minimum number of nucleotide differences between individual barcodes. Based on standard error rates for pyrosequencing and our observation that ~6% of sequences have at least one error in either their forward or reverse 10-nt barcodes (Table 3, and previous paragraph), we can predict that if the minimum number of nucleotide differences is decreased, there would be an increase in the frequency of sequence misassignment.

Our barcode specification has been tailored for use with pyrosequencing technologies, particularly reflecting restrictions imposed on the generation of homopolymeric runs. These barcodes may also be compatible with other sequencing platforms. Optimally, however the precise structure of the barcodes might need to be revised to reflect sequencing error profiles of the various systems. If necessary, multiple sequencing primers may also be used to identify the barcode and the sequence of interest, to overcome limitations imposed by short read lengths. As the availability of high-throughput technology advances, massively multiplexed barcoded sequencing will be extremely useful in performing surveys of diverse systems for nucleic acid signatures.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was funded by Stanford University Department of Pathology, NIH ROI GM37706; SGTC: 1S10RR022982 and R01HG003571. We wish to thank Jonathan Gent, Scott Boyd, Julia Pak, Weng-Onn Lui,

Sam Gu, Daniel Blanchard, Rosa Alcazar, Frederick Tan, Ky Sha, Steve Johnson, Jay Maniar, Cecilia Mello and Chaya Krishna for assistance with materials, technical aspects and discussion. P.P. is supported by a Stanford Graduate Fellowship. Funding to pay the Open Access publication charges for this article was provided by Stanford University Pathology Department.

Conflict of interest statement. None declared.

REFERENCES

- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373–382.
- Ahmadian, A., Ehn, M. and Hober, S. (2006) Pyrosequencing: history, biochemistry and future. *Clin. Chim. Acta*, **363**, 83–94.
- Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T. *et al.* (2006) Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect. Immun.*, **74**, 4694–4707.
- Oh, J.D., Kling-Backhed, H., Giannakis, M., Xu, J., Fulton, R.S., Fulton, L.A., Cordum, H.S., Wang, C., Elliott, G. *et al.* (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *Proc. Natl Acad. Sci. USA*, **103**, 9999–10004.
- Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A. and Carrington, J.C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol.*, **5**, e57.
- Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, **127**, 1193–1207.
- Thomas, R.K., Nickerson, E., Simons, J.F., Janne, P.A., Tengs, T., Yuza, Y., Garraway, L.A., LaFramboise, T., Lee, J.C. *et al.* (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med.*, **12**, 852–855.
- Mazurkiewicz, P., Tang, C.M., Boone, C. and Holden, D.W. (2006) Signature-tagged mutagenesis: barcoding mutants for genome-wide screens. *Nat. Rev. Genet.*, **7**, 929–939.
- Binladen, J., Gilbert, M.T., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R. and Willerslev, E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.
- Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M.Q., Tebas, P. and Bushman, F.D. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.*, **35**, e91.
- Lau, N.C., Lim, L.P., Weinstein, E.G. and Bartel, D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
- Pak, J. and Fire, A. (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science*, **315**, 241–244.

Supplementary Table 1: Complete list of Barcoded Primers. 5P: 5'-P-dependent cloning; DL: 5'-P-independent cloning. Primers are of the type Adapter::Barcode::Cloning-Linker. Unique sets of forward and reverse primers were used for each sample. **1A:** Primers used in run I. **1B:** Primers used in run II.

Supplementary Table 2: Sanger sequencing data for the two pyrosequencing runs. 5P: 5'-P-dependent cloning; DL: 5'-P-independent cloning. All samples are from virus-infected host systems. In both runs, all 48 clones that were sequenced had unambiguously identifiable barcodes at both flanking regions. 29.16% and 29.17% of the PCR products from runs I and II respectively, did not have small RNA inserts. There appeared to be no contaminants.

Supplementary Table 3: Representation of the types of errors encountered in barcode motifs. From each sequencing run, two hundred five sequences that had one perfect 13 nt. motif, but contained errors in the second barcode motif, were arbitrarily chosen for analyzing the number of indels and substitutions associated with homo-polymeric runs (in the barcode motif with errors).

Supplementary Tables 4A – 4F (Amplicon Run I), and 5A – 5F (Amplicon Run II): False-discovery rate as a function of search motif length. The term 'search motif' refers to barcode plus cloning-linker sequence that may be used for distinguishing sequences from various independent experiments. Only sequences that perfectly match the search motif are assigned to independent bins. The output from Barverify is displayed in a tabular format for each search motif. Designs for the search motifs include the complete barcode, but vary in the length of

contributing sequence from cloning-linkers (0, 1, 2, 3 or 4 nucleotides, or complete sequence). The contributing sequence-lengths from the barcode and from the cloning-linker adjacent to the barcode are indicated in parentheses in the titles for the tables. Rows denote search motifs that probe at the 5' end of the sequencing read, and columns denote search motifs that probe at the 3' end of the sequencing read. Number of sequences that contain identifiable, perfectly matched 5' and 3' search motifs is represented for every combination of 5' and 3' motifs. Sequences that have a perfect 5' motif, but have imperfect 3' motifs are represented in the column marked “(close)” if they pass a minimum score-threshold for identity to expected 3' motif; else they are represented in the column marked “(no match).” Sequences that have a perfect 3' motif, but have imperfect 5' motifs are represented in the row marked “(close)” if they pass a minimum score-threshold for identity to expected 5' motif; else they are represented in the row marked “(no match).” Score for the match is calculated by: $(2 * \text{number of correct matches} - \text{number of insertions and deletions}) / (2 * \text{length of expected primer})$. For a sequence to have a “close match,” its score should be greater than 0.88. Analysis of data from both rounds of pyrosequencing indicates that the number of false-assigned sequences (i.e. number of sequences that are in off-diagonal squares, excluding the number of imperfect matches) progressively decreases with an increase in search stringency, i.e. by increasing length of the search motif.

SUPPLEMENTARY METHODS:

Barsort is a program that sorts a series of barcoded reads from the sequencer into bins according to their barcodes or according to specified search motifs. The program takes two input files: a list of experiments with their specific 5' and 3' search motifs, and the set of raw reads from the sequencer given in FASTA format. The program assigns each sequence to an experiment based

on whether it contains an exact match to the 5' and/or 3' motifs. Sequences that contain both 5' and 3' motifs are given first priority, followed by sequences that contain either the 5' or the 3' motif. Sequences that do not contain perfect matches to any motifs are put into an "unassigned" bin for further analysis. Sequences that have been recognized to contain either or both motifs are FASTA-formatted, and placed in their experiment's "raw" bin. Additionally, these sequences are stripped of their barcode motifs, and placed in their experiment's "trimmed" bin. The files in the "trimmed" bin are in FASTA format.

Barverify is a program that analyzes the output from Barsort to identify potential sources of contamination. The program looks at the "raw" files generated by Barsort, and scans for reads that contain only 5' or 3' search motifs. With the identified motif anchored, the program tries to find a match for the missing motif from the entered set of all experimental motifs, and selects the best-scored match based on alignment of individual motifs with input sequence. The score is calculated in terms of mismatches, insertions, and deletions relative to the length of the search motif. The results of the scans and alignments are provided in the "barcode matches" file for each barcoded set of sequences. For each 5' and 3' barcode motif, the program also generates numeric maps comparing how often perfect 5'-3', 5'-5', and 3'-3' barcode motif pairs are encountered in the entire sequence set. Imperfect matches are grouped into two columns based on: (a) best match to the expected member of the barcode-pair (with a score above the set threshold), or (b) a low-scoring match to the expected barcode, or to a barcode from another sample set.

SUPPLEMENTARY DISCUSSION:

Evaluating the use of Sanger sequencing for quality control before pyrosequencing

We assessed data from Sanger-sequenced clones, to evaluate the use of small-scale conventional sequencing for quality control purposes. The Sanger-sequenced data set was by no means exhaustive (number of clones sequenced was only 48), was prone to inconsistencies introduced by random sampling error, and was limited to molecules that were double-stranded (which in turn restricts material to molecules that have amplified in the last round of PCR, and to pairs of abundant molecules that can reanneal easily). Despite these limitations, we observed that for both runs (using any search motif length), data from Sanger sequencing [Supplementary Table 2] and from pyrosequencing [Supplementary Tables 4, 5] were similar with respect to barcode distribution across individual libraries, and to the nature of small RNA insert sequences. 100% of the sequences in both Sanger runs, and 99.87% of sequences in Run I and 99.83% of sequences in Run II had at least one barcode motif (10 nt. barcode + 3 nt. adjoining cloning-linker) with no errors. This correlation between the data sets establishes Sanger sequencing as a means to partially verify sequence distribution and quality, and to identify possible contaminants, if they were highly amplified.

Error Analysis

It was observed that with a 13 nt. barcode motif, 70% - 85% of sequences with imperfect search motifs were highly identical to the expected search motif. Further analysis of the sequences that remained unassigned demonstrated that 70% of them had errors were in the barcode, and 30% of them had errors in the 3 nucleotides of the motif corresponding to Cloning-Linker sequence.

SUPPLEMENTARY TABLE 1A

SAMPLE	METHOD	REV PRIMER NAME	REV PRIMER SEQUENCE	FOR PRIMER NAME	FOR PRIMER SEQUENCE
454-41	5P	AF-PP-165	GCCTCCCTCGCGCCA-TCAG-AGCCTAAGCT-ATCGTAGGCACCTGAAA	AF-PP-166	GCCTTGCCAGCCCCGC-TCAG-AGCTTAGGCT-ACCGTCCTTGGTGCCCCG
454-42	DL	AF-PP-217	GCCTCCCTCGCGCCA-TCAG-AGTTCAAGTC-TCCCTGGCACCCGAGTG	AF-PP-229	GCCTTGCCAGCCCCGC-TCAG-AGTCCAGGTC-TTGATGGTGCCTACAG
454-43	5P	AF-PP-159; AF-PP-161	GCCTCCCTCGCGCCA-TCAG-ACTTGAAGTG-ATCGTAGGCACCTGAAA A	AF-PP-160; AF-PP-162	GCCTTGCCAGCCCCGC-TCAG-ACTGGACCTG-ACCGTCCTTGGTGCCCCG
454-44	DL	AF-PP-218	GCCTCCCTCGCGCCA-TCAG-ACGGTAACGT-TCCCTGGCACCCGAGTG	AF-PP-230	GCCTTGCCAGCCCCGC-TCAG-ACGTTACCGT-TTGATGGTGCCTACAG
454-45	5P	AF-PP-173	GCCTCCCTCGCGCCA-TCAG-ATCCGAATCG-ATCGTAGGCACCTGAAA	AF-PP-174	GCCTTGCCAGCCCCGC-TCAG-ATCGGATTTCG-ACCGTCCTTGGTGCCCCG
454-46	DL	AF-PP-219	GCCTCCCTCGCGCCA-TCAG-ATGGCAATGC-TCCCTGGCACCCGAGTG	AF-PP-231	GCCTTGCCAGCCCCGC-TCAG-ATGCCATTGC-TTGATGGTGCCTACAG
454-47	5P	AF-PP-177	GCCTCCCTCGCGCCA-TCAG-CAGGTCCAGT-ATCGTAGGCACCTGAAA	AF-PP-178	GCCTTGCCAGCCCCGC-TCAG-CAGTTCAAGT-ACCGTCCTTGGTGCCCCG
454-48	DL	AF-PP-220	GCCTCCCTCGCGCCA-TCAG-CATTGCCATG-TCCCTGGCACCCGAGTG	AF-PP-232	GCCTTGCCAGCCCCGC-TCAG-CATGGCAATG-TTGATGGTGCCTACAG
454-49	5P	AF-PP-181	GCCTCCCTCGCGCCA-TCAG-CTAAGCCTAG-ATCGTAGGCACCTGAAA	AF-PP-182	GCCTTGCCAGCCCCGC-TCAG-CTAGGCTTAG-ACCGTCCTTGGTGCCCCG
454-50	DL	AF-PP-221	GCCTCCCTCGCGCCA-TCAG-CGAATCCGAT-TCCCTGGCACCCGAGTG	AF-PP-222	GCCTTGCCAGCCCCGC-TCAG-CGATTCTGGAT-TTGATGGTGCCTACAG
454-51	5P	AF-PP-185	GCCTCCCTCGCGCCA-TCAG-TCAAGTTAGC-ATCGTAGGCACCTGAAA	AF-PP-186	GCCTTGCCAGCCCCGC-TCAG-TAGCCTAAGC-ACCGTCCTTGGTGCCCCG
454-52	DL	AF-PP-223	GCCTCCCTCGCGCCA-TCAG-TACCGTTACG-TCCCTGGCACCCGAGTG	AF-PP-224	GCCTTGCCAGCCCCGC-TCAG-TACGGTAACG-TTGATGGTGCCTACAG
454-53	5P	AF-PP-189	GCCTCCCTCGCGCCA-TCAG-TGAAGTTGAC-ATCGTAGGCACCTGAAA	AF-PP-190	GCCTTGCCAGCCCCGC-TCAG-TGACCTGGAC-ACCGTCCTTGGTGCCCCG
454-54	DL	AF-PP-225	GCCTCCCTCGCGCCA-TCAG-TAGGCTTCAG-TCCCTGGCACCCGAGTG	AF-PP-226	GCCTTGCCAGCCCCGC-TCAG-TCAGGTCCAG-TTGATGGTGCCTACAG
454-55	5P	AF-PP-193	GCCTCCCTCGCGCCA-TCAG-AGCCTCCAGT-ATCGTAGGCACCTGAAA	AF-PP-194	GCCTTGCCAGCCCCGC-TCAG-AGCTTCTTAG-ACCGTCCTTGGTGCCCCG
454-56	DL	AF-PP-227	GCCTCCCTCGCGCCA-TCAG-AGCCTGGCAT-TCCCTGGCACCCGAGTG	AF-PP-233	GCCTTGCCAGCCCCGC-TCAG-AGCTTGCCAT-TTGATGGTGCCTACAG
454-57	5P	AF-PP-197	GCCTCCCTCGCGCCA-TCAG-AGTTCGGACT-	AF-PP-198	GCCTTGCCAGCCCCGC-TCAG-AGTCCGAAGT-

			ATCGTAGGCACCTGAAA		ACCGTCCTTGGTGCCCG
454-58	DL	AF-PP-228	GCCTCCCTCGCGCCA-TCAG-AGTTCTTGAC-TCCTTGGCACCCGAGTG	AF-PP-234	GCCTTGCCAGCCCGC-TCAG-AGTCCTCCAG-TTGATGGTGCCTACAG
454-59	5P	AF-PP-201	GCCTCCCTCGCGCCA-TCAG-ACGGTCCATG-ATCGTAGGCACCTGAAA	AF-PP-202	GCCTTGCCAGCCCGC-TCAG-ACGTTCAATG-ACCGTCCTTGGTGCCCG
454-60	5P	AF-PP-203	GCCTCCCTCGCGCCA-TCAG-ACTTGTTTCAG-ATCGTAGGCACCTGAAA	AF-PP-204	GCCTTGCCAGCCCGC-TCAG-ATGGTGGAC-ACCGTCCTTGGTGCCCG
454-61	5P	AF-PP-205	GCCTCCCTCGCGCCA-TCAG-ACTTGCCGAT-ATCGTAGGCACCTGAAA	AF-PP-206	GCCTTGCCAGCCCGC-TCAG-ACTGGCGGAT-ACCGTCCTTGGTGCCCG
454-62	5P	AF-PP-207	GCCTCCCTCGCGCCA-TCAG-ACGGTGGATC-ATCGTAGGCACCTGAAA	AF-PP-208	GCCTTGCCAGCCCGC-TCAG-ACGTTGAATC-ACCGTCCTTGGTGCCCG
454-63	5P	AF-PP-209	GCCTCCCTCGCGCCA-TCAG-ATCCGCCTAG-ATCGTAGGCACCTGAAA	AF-PP-210	GCCTTGCCAGCCCGC-TCAG-ATCGGCAAGT-ACCGTCCTTGGTGCCCG
454-64	5P	AF-PP-211	GCCTCCCTCGCGCCA-TCAG-ATCCGTTAGC-ATCGTAGGCACCTGAAA	AF-PP-212	GCCTTGCCAGCCCGC-TCAG-ATCGGTAAGC-ACCGTCCTTGGTGCCCG
454-65	5P	AF-PP-213	GCCTCCCTCGCGCCA-TCAG-ATGGCGGTAC-ATCGTAGGCACCTGAAA	AF-PP-214	GCCTTGCCAGCCCGC-TCAG-ATGCCGTTAC-ACCGTCCTTGGTGCCCG

SUPPLEMENTARY TABLE 1B

SAMPLE	METHOD	REV PRIMER NAME	REV PRIMER SEQUENCE	FOR PRIMER NAME	FOR PRIMER SEQUENCE
454-66	5P	AF-PP-165	GCCTCCCTCGCGCCA-TCAG-AGCCTAAGCT-ATCGTAGGCACCTGAAA	AF-PP-230	GCCTTGCCAGCCCGC-TCAG-ACGTTACCGT-TTGATGGTGCCTACAG
454-67	5P	AF-PP-167	GCCTCCCTCGCGCCA-TCAG-AGTTCAAGTC-ATCGTAGGCACCTGAAA	AF-PP-168	GCCTTGCCAGCCCGC-TCAG-AGTCCAGGTC-ACCGTCCTTGGTGCCCG
454-68	5P	AF-PP-159; AF-PP-161	GCCTCCCTCGCGCCA-TCAG-ACTTGAAGTG-ATCGTAGGCACCTGAAA	AF-PP-160; AF-PP-162	GCCTTGCCAGCCCGC-TCAG-ATCTGGACCTG-ACCGTCCTTGGTGCCCG
454-69	5P	AF-PP-171	GCCTCCCTCGCGCCA-TCAG-ACGGTAACGT-ATCGTAGGCACCTGAAA	AF-PP-166	GCCTTGCCAGCCCGC-TCAG-AGCTTAGGCT-ACCGTCCTTGGTGCCCG
454-70	5P	AF-PP-173	GCCTCCCTCGCGCCA-TCAG-ATCCGAATCG-ATCGTAGGCACCTGAAA	AF-PP-174	GCCTTGCCAGCCCGC-TCAG-ATCGGATTTCG-ACCGTCCTTGGTGCCCG
454-71	5P	AF-PP-175	GCCTCCCTCGCGCCA-TCAG-ATGGCAATGC-ATCGTAGGCACCTGAAA	AF-PP-176	GCCTTGCCAGCCCGC-TCAG-ATGCCATTGC-ACCGTCCTTGGTGCCCG
454-72	5P	AF-PP-177	GCCTCCCTCGCGCCA-TCAG-CAGGTCCAGT-ATCGTAGGCACCTGAAA	AF-PP-178	GCCTTGCCAGCCCGC-TCAG-CAGTTCAAGT-ACCGTCCTTGGTGCCCG
454-73	5P	AF-PP-179	GCCTCCCTCGCGCCA-TCAG-CATTGCCATG-ATCGTAGGCACCTGAAA	AF-PP-198	GCCTTGCCAGCCCGC-TCAG-AGTCCGAAGT-ACCGTCCTTGGTGCCCG
454-74	5P	AF-PP-181	GCCTCCCTCGCGCCA-TCAG-CTAAGCCTAG-ATCGTAGGCACCTGAAA	AF-PP-180	GCCTTGCCAGCCCGC-TCAG-CATGGCAATG-ACCGTCCTTGGTGCCCG
454-75	5P	AF-PP-183	GCCTCCCTCGCGCCA-TCAG-CGAATCCGAT-ATCGTAGGCACCTGAAA	AF-PP-184	GCCTTGCCAGCCCGC-TCAG-CGATTTCGGAT-ACCGTCCTTGGTGCCCG
454-76	5P	AF-PP-185	GCCTCCCTCGCGCCA-TCAG-TCAAGTTAGC-ATCGTAGGCACCTGAAA	AF-PP-186	GCCTTGCCAGCCCGC-TCAG-TAGCCTAAGC-ACCGTCCTTGGTGCCCG
454-77	5P	AF-PP-187	GCCTCCCTCGCGCCA-TCAG-TACCGTTACG-ATCGTAGGCACCTGAAA	AF-PP-188	GCCTTGCCAGCCCGC-TCAG-TACGGTAACG-ACCGTCCTTGGTGCCCG
454-78	5P	AF-PP-189	GCCTCCCTCGCGCCA-TCAG-TGAAGTTGAC-ATCGTAGGCACCTGAAA	AF-PP-190	GCCTTGCCAGCCCGC-TCAG-TGACCTGGAC-ACCGTCCTTGGTGCCCG
454-79	5P	AF-PP-191	GCCTCCCTCGCGCCA-TCAG-TAGGCTTCAG-ATCGTAGGCACCTGAAA	AF-PP-192	GCCTTGCCAGCCCGC-TCAG-TCAGGTCCAG-ACCGTCCTTGGTGCCCG
454-80	5P	AF-PP-193	GCCTCCCTCGCGCCA-TCAG-AGCCTCCAGT-ATCGTAGGCACCTGAAA	AF-PP-194	GCCTTGCCAGCCCGC-TCAG-AGCTTCTTAG-ACCGTCCTTGGTGCCCG
454-81	5P	AF-PP-195	GCCTCCCTCGCGCCA-TCAG-AGCCTGGCAT-ATCGTAGGCACCTGAAA	AF-PP-196	GCCTTGCCAGCCCGC-TCAG-AGCTTGCCAT-ACCGTCCTTGGTGCCCG
454-82	5P	AF-PP-197	GCCTCCCTCGCGCCA-TCAG-AGTTCGGACT-ATCGTAGGCACCTGAAA	AF-PP-182	GCCTTGCCAGCCCGC-TCAG-CTAGGCTTAG-ACCGTCCTTGGTGCCCG

454-83	5P	AF-PP-199	GCCTCCCTCGCGCCA-TCAG-AGTTCTTGAC-ATCGTAGGCACCTGAAA	AF-PP-200	GCCTTGCCAGCCCGC-TCAG-AGTCCTCCAG-ACCGTCCTTGGTGCCCG
454-84	5P	AF-PP-201	GCCTCCCTCGCGCCA-TCAG-ACGGTCCATG-ATCGTAGGCACCTGAAA	AF-PP-202	GCCTTGCCAGCCCGC-TCAG-ACGTTCAATG-ACCGTCCTTGGTGCCCG
454-85	5P	AF-PP-203	GCCTCCCTCGCGCCA-TCAG-ACTTGTTTCAG-ATCGTAGGCACCTGAAA	AF-PP-204	GCCTTGCCAGCCCGC-TCAG-ACTGGTGGAC-ACCGTCCTTGGTGCCCG
454-86	5P	AF-PP-205	GCCTCCCTCGCGCCA-TCAG-ACTTGCCGAT-ATCGTAGGCACCTGAAA	AF-PP-206	GCCTTGCCAGCCCGC-TCAG-ACTGGCGGAT-ACCGTCCTTGGTGCCCG
454-87	5P	AF-PP-207	GCCTCCCTCGCGCCA-TCAG-ACGGTGGATC-ATCGTAGGCACCTGAAA	AF-PP-208	GCCTTGCCAGCCCGC-TCAG-ACGTTGAATC-ACCGTCCTTGGTGCCCG
454-88	5P	AF-PP-209	GCCTCCCTCGCGCCA-TCAG-ATCCGCCTAG-ATCGTAGGCACCTGAAA	AF-PP-210	GCCTTGCCAGCCCGC-TCAG-ATCGGCAAGT-ACCGTCCTTGGTGCCCG
454-89	5P	AF-PP-211	GCCTCCCTCGCGCCA-TCAG-ATCCGTTAGC-ATCGTAGGCACCTGAAA	AF-PP-212	GCCTTGCCAGCCCGC-TCAG-ATCGGTAAGC-ACCGTCCTTGGTGCCCG
454-90	5P	AF-PP-213	GCCTCCCTCGCGCCA-TCAG-ATGGCGGTAC-ATCGTAGGCACCTGAAA	AF-PP-214	GCCTTGCCAGCCCGC-TCAG-ATGCCGTTAC-ACCGTCCTTGGTGCCCG

SUPPLEMENTARY TABLE 2

Name Assigned to Sample: Run I	Cloning Method: Run I	# of Clones: Run I
454-41	5P	3
454-42	DL	3
454-43	5P	3
454-44	DL	4
454-45	5P	1
454-46	DL	1
454-47	5P	2
454-48	DL	0
454-49	5P	2
454-50	DL	2
454-51	5P	0
454-52	DL	2
454-53	5P	4
454-54	DL	2
454-55	5P	3
454-56	DL	4
454-57	5P	3
454-58	DL	1
454-59	5P	3
454-60	5P	2
454-61	5P	0
454-62	5P	1
454-63	5P	2
454-64	5P	0
454-65	5P	0

Name Assigned to Sample: Run II	Cloning Method: Run II	# of Clones: Run II
454-66	5P	2
454-67	5P	1
454-68	5P	3
454-69	5P	1
454-70	5P	4
454-71	5P	1
454-72	5P	3
454-73	5P	1
454-74	5P	3
454-75	5P	1
454-76	5P	1
454-77	5P	1
454-78	5P	1
454-79	5P	2
454-80	5P	6
454-81	5P	1
454-82	5P	0
454-83	5P	2
454-84	5P	3
454-85	5P	2
454-86	5P	4
454-87	5P	1
454-88	5P	1
454-89	5P	3
454-90	5P	0

SUPPLEMENTARY TABLE 3

ERROR CATEGORY	REPRESENTATIVE # IN RUN I	REPRESENTATIVE # IN RUN II
Deletion of a nucleotide that is NOT part of a homo-polymeric run	42	54
Deletion of a nucleotide that is part of a homo-polymeric run	21	16
Insertion of a nucleotide NOT in the context of a homo-polymeric run	22	9
Insertion of a nucleotide in the context of a homo-polymeric run	8	18
Insertion of a nucleotide NOT in the context of an existing homo-polymeric run, but one that generates a homo-polymeric run	18	8
Substitution of a nucleotide NOT in the context of a homo-polymeric run	13	12
Substitution of a nucleotide in the context of a homo-polymeric run	27	57
Substitution of a nucleotide NOT in the context of an existing homo-polymeric run, but one that generates a homo-polymeric run	44	28
Multiple errors	7	1
Presence of N's	3	2

TOTAL

205

205

SUPPLEMENTARY TABLES 4A – 4F: AMPLICON RUN I

A. SEARCH MOTIF = Barcode (10 nt)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	8953	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-42	0	10469	14	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	125
454-43	0	0	13991	0	0	0	450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-44	0	0	0	10610	0	0	0	0	0	0	0	126	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-45	0	0	0	0	7797	0	0	0	0	512	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-46	0	0	0	0	0	6681	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	139
454-47	0	0	444	0	0	0	7327	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-48	0	0	0	0	0	0	0	7910	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-49	0	0	0	0	0	0	0	0	7695	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-50	0	0	0	0	405	0	0	0	0	11780	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-51	0	0	0	0	0	0	0	0	0	0	9781	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61	164
454-52	0	0	0	7	0	0	0	0	0	0	0	17465	0	0	0	0	0	0	0	0	0	0	0	0	0	0	230
454-53	0	0	0	0	0	0	0	0	0	0	0	0	9231	0	0	0	0	0	0	0	0	0	0	0	0	65	101
454-54	0	0	10	0	0	0	0	0	0	0	0	0	0	9055	0	0	0	0	0	0	0	0	0	0	0	8	127
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8090	0	0	0	0	0	0	0	0	0	0	201	133
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9423	0	0	0	0	0	0	0	0	0	65	94
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10909	0	0	0	0	0	0	0	0	0	0
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9358	0	0	0	0	0	0	0	54	112
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9180	0	0	0	0	0	0	192	124
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	10569	0	0	0	0	0	87	136
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8159	0	378	0	0	0	0
454-62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10476	0	0	0	209	152
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7876	0	0	0	163
454-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2825	1	16	48
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3667	32	42
(close)	0	30	0	0	0	39	0	0	0	0	35	0	17	0	18	66	0	0	105	133	1	120	0	19	15		
(no match)	0	170	0	0	0	140	0	0	0	0	131	0	162	0	119	108	0	300	89	178	148	107	0	41	32		

B. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (1 nt)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	8503	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136	87
454-42	0	10438	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	127
454-43	0	0	13961	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	254
454-44	2	0	0	10278	0	0	0	0	0	0	0	224	0	0	0	0	0	0	0	0	0	0	0	0	0	139	65
454-45	0	0	0	0	7772	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102	154
454-46	0	0	0	0	0	6666	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76	71
454-47	0	0	0	0	0	0	7305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	175
454-48	0	0	0	0	0	0	0	7540	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	60
454-49	0	0	0	0	0	0	0	0	7334	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	74
454-50	0	0	0	0	0	0	0	0	0	11738	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	165
454-51	0	0	0	0	0	0	0	0	0	15	9757	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85	154
454-52	0	0	0	0	0	0	0	0	0	0	0	17447	0	0	0	0	0	0	0	0	0	0	0	0	0	12	225
454-53	0	0	0	0	0	0	0	0	0	1	0	0	9195	0	0	0	0	0	0	0	0	0	0	0	0	80	107
454-54	0	0	2	0	0	0	0	0	0	0	0	0	0	9041	0	0	0	0	0	0	0	0	0	0	0	0	145
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8066	0	0	0	0	0	0	0	0	0	0	211	134
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9400	0	0	0	0	0	0	0	0	0	69	95
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10430	0	0	0	0	0	0	0	0	154	86
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9338	0	0	0	0	0	0	0	73	108
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9157	0	0	0	0	0	0	208	120
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10514	0	0	0	0	0	97	159
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8131	0	0	0	0	47	161
454-62	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10444	0	0	0	214	162
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7852	0	0	59	112
454-64	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	2816	1	17	52
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3662	23	53
(close)	38	67	0	18	68	45	75	9	55	91	36	0	81	0	57	66	92	203	111	169	57	112	83	25	17		
(no match)	188	153	0	109	125	146	144	183	104	194	140	0	113	0	93	126	146	110	94	165	112	133	114	39	33		

C. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (2 nt)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	8490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	81	148
454-42	0	10390	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	155
454-43	0	0	13925	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	167	259
454-44	2	0	0	10209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	162
454-45	0	0	0	0	7745	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	115	145
454-46	0	0	0	0	0	6644	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	64
454-47	0	0	0	0	0	0	7282	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	140
454-48	0	0	0	0	0	0	0	7514	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	130
454-49	0	0	0	0	0	0	0	0	7318	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	120
454-50	0	0	0	0	0	0	0	0	0	11673	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	133	144
454-51	0	0	0	0	0	0	0	0	0	0	9746	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101	143
454-52	0	0	0	1	0	0	0	0	0	0	0	17401	0	0	0	0	0	0	0	0	0	0	0	0	0	91	192
454-53	0	0	0	0	0	0	0	0	0	0	0	0	9175	0	0	0	0	0	0	0	0	0	0	0	0	81	113
454-54	0	0	0	0	0	0	0	0	0	0	0	0	0	9014	0	0	0	0	0	0	0	0	0	0	0	45	122
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8054	0	0	0	0	0	0	0	0	0	0	217	134
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9353	0	0	0	0	0	0	0	0	0	107	93
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10405	0	0	0	0	0	0	0	0	102	150
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9312	0	0	0	0	0	0	0	78	119
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9138	0	0	0	0	0	0	216	121
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10487	0	0	0	0	0	98	165
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8108	0	0	0	0	89	131
454-62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10421	0	0	0	223	163
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7837	0	0	60	112
454-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2811	1	19	51
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3656	25	53
(close)	100	71	101	79	73	53	103	65	62	136	73	33	83	27	58	82	93	203	113	175	62	114	81	28	20		
(no match)	133	161	179	63	146	147	130	151	107	185	109	204	124	195	99	121	159	120	102	179	118	144	130	40	34		

D. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (3 nt)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	8460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	178	69
454-42	0	10338	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	190	51
454-43	0	0	13873	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	389	71
454-44	2	0	0	10013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	387	47
454-45	0	0	0	0	7704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	175	109
454-46	0	0	0	0	0	6609	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	130	42
454-47	0	0	0	0	0	0	7243	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	177	78
454-48	0	0	0	0	0	0	0	7471	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	167	39
454-49	0	0	0	0	0	0	0	0	7286	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	192	33
454-50	0	0	0	0	0	0	0	0	0	11510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	328	86
454-51	0	0	0	0	0	0	0	0	0	0	9709	0	0	0	0	0	0	0	0	0	0	0	0	0	0	200	69
454-52	0	0	0	0	0	0	0	0	0	0	0	17261	0	0	0	0	0	0	0	0	0	0	0	0	0	341	55
454-53	0	0	0	0	0	0	0	0	0	0	0	0	9140	0	0	0	0	0	0	0	0	0	0	0	0	159	56
454-54	0	0	0	0	0	0	0	0	0	0	0	0	0	8963	0	0	0	0	0	0	0	0	0	0	0	160	34
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8025	0	0	0	0	0	0	0	0	0	0	261	107
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9211	0	0	0	0	0	0	0	0	0	278	52
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10357	0	0	0	0	0	0	0	0	212	64
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9238	0	0	0	0	0	0	0	186	61
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9099	0	0	0	0	0	0	318	44
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10428	0	0	0	0	0	228	74
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8068	0	0	0	0	177	69
454-62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10379	0	0	0	357	55
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7806	0	0	148	35
454-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2804	1	45	28
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3650	57	24
(close)	185	185	226	125	158	126	112	112	132	194	136	225	174	168	119	131	219	275	198	296	142	219	187	44	35		
(no match)	60	73	72	37	78	97	136	121	52	153	58	40	47	80	50	84	57	72	31	78	52	55	44	28	22		

E. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (4 nt)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	8412	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	196	70
454-42	0	10307	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	210	51
454-43	0	0	13829	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	414	72
454-44	2	0	0	9957	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	448	29
454-45	0	0	0	0	7666	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	177	119
454-46	0	0	0	0	0	6586	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	134	46
454-47	0	0	0	0	0	0	7207	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	181	85
454-48	0	0	0	0	0	0	0	7448	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	202	24
454-49	0	0	0	0	0	0	0	0	7267	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	199	33
454-50	0	0	0	0	0	0	0	0	0	11454	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	392	63
454-51	0	0	0	0	0	0	0	0	0	0	9673	0	0	0	0	0	0	0	0	0	0	0	0	0	0	214	70
454-52	0	0	0	0	0	0	0	0	0	0	0	17221	0	0	0	0	0	0	0	0	0	0	0	0	0	366	52
454-53	0	0	0	0	0	0	0	0	0	0	0	0	9109	0	0	0	0	0	0	0	0	0	0	0	0	170	58
454-54	0	0	0	0	0	0	0	0	0	0	0	0	0	8934	0	0	0	0	0	0	0	0	0	0	0	144	60
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7986	0	0	0	0	0	0	0	0	0	0	267	120
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9164	0	0	0	0	0	0	0	0	0	323	39
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10298	0	0	0	0	0	0	0	0	236	66
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9212	0	0	0	0	0	0	0	214	48
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9070	0	0	0	0	0	0	326	44
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10375	0	0	0	0	0	247	75
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8042	0	0	0	0	186	68
454-62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10335	0	0	0	375	53
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7766	0	0	166	35
454-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2798	1	45	29
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3628	69	25
(close)	215	201	244	153	179	153	128	167	146	259	158	245	191	187	139	167	253	309	219	332	160	250	211	49	44		
(no match)	59	68	72	22	83	85	145	69	50	103	57	38	48	80	50	63	56	49	31	75	52	52	42	28	22		

F. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (complete sequence)

	454-41	454-42	454-43	454-44	454-45	454-46	454-47	454-48	454-49	454-50	454-51	454-52	454-53	454-54	454-55	454-56	454-57	454-58	454-59	454-60	454-61	454-62	454-63	454-64	454-65	(close)	(no match)
454-41	7811	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	652	28
454-42	0	9802	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	403	49
454-43	0	0	12605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1342	37
454-44	1	0	0	9508	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	650	16
454-45	0	0	0	0	7015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	710	13
454-46	0	0	0	0	0	6197	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	284	11
454-47	0	0	0	0	0	0	6634	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	597	50
454-48	0	0	0	0	0	0	0	7045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	372	25
454-49	0	0	0	0	0	0	0	0	6616	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	694	29
454-50	0	0	0	0	0	0	0	0	0	10823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	670	29
454-51	0	0	0	0	0	0	0	0	0	0	9196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	540	64
454-52	0	0	0	0	0	0	0	0	0	0	0	16521	0	0	0	0	0	0	0	0	0	0	0	0	0	655	35
454-53	0	0	0	0	0	0	0	0	0	0	0	0	7522	0	0	0	0	0	0	0	0	0	0	0	0	1564	30
454-54	0	0	0	0	0	0	0	0	0	0	0	0	0	8389	0	0	0	0	0	0	0	0	0	0	0	451	22
454-55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7459	0	0	0	0	0	0	0	0	0	0	677	61
454-56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8671	0	0	0	0	0	0	0	0	0	547	21
454-57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9403	0	0	0	0	0	0	0	0	965	24
454-58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8759	0	0	0	0	0	0	0	409	56
454-59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8273	0	0	0	0	0	0	907	32
454-60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9253	0	0	0	0	0	1117	42
454-61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5969	0	0	0	0	2097	41
454-62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9403	0	0	0	1080	40
454-63	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6956	0	0	788	27
454-64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2245	0	511	56
454-65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2945	692	12
(close)	429	534	612	424	449	459	435	411	324	653	328	649	436	476	333	474	479	581	432	605	371	495	402	125	135		
(no match)	32	49	35	12	37	53	30	57	32	96	44	62	24	67	32	43	38	27	46	87	30	47	47	13	4		

SUPPLEMENTARY TABLES 5A – 5F: AMPLICON RUN II

A. SEARCH MOTIF = Barcode (10 nt)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13918	0	0	569	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-67	0	7549	22	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	13	177
454-68	0	0	13388	0	0	0	474	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-69	541	0	0	6582	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-70	0	0	0	0	9594	0	0	0	0	558	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-71	0	0	0	0	0	9498	0	11	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	63	224
454-72	0	0	455	0	0	0	12472	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-73	0	0	0	0	0	0	0	7349	335	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-74	0	0	0	0	0	0	0	0	14028	0	0	0	0	0	0	0	496	0	0	0	0	0	0	0	0	0	0
454-75	0	0	0	0	609	0	0	0	0	15262	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
454-76	0	0	0	0	0	0	0	0	0	0	10396	0	0	0	0	0	0	0	0	0	0	0	0	0	0	89	194
454-77	1	1	0	0	0	0	0	0	0	0	0	9710	0	0	0	0	0	0	0	0	0	0	0	0	0	85	141
454-78	0	0	0	0	0	0	0	0	0	0	0	0	12219	0	0	0	0	0	0	0	0	0	0	0	0	101	94
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	10470	0	0	0	0	0	0	0	0	0	0	0	114	133
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10160	0	0	0	0	2	0	0	0	0	0	156	208
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6881	0	0	0	0	0	0	0	0	0	90	114
454-82	0	0	0	0	0	0	0	450	0	0	0	0	0	0	0	0	11839	0	0	0	0	0	0	0	0	0	0
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12283	0	0	0	0	0	0	0	155	118
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14181	0	0	0	0	0	0	231	193
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7724	0	0	0	0	0	59	89
454-86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8978	0	517	0	0	0	0
454-87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10585	0	0	0	208	149
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10505	0	0	0	217
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8257	1	71	131
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5101	55	64
(close)	0	0	0	0	0	55	0	0	0	0	33	22	31	27	35	102	0	135	165	119	1	149	0	89	64		
(no match)	0	0	0	0	0	82	0	0	0	0	163	183	274	300	205	121	0	88	159	83	201	83	0	107	62		

B. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (1 nt)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13852	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	188	298
454-67	0	7520	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	83	141
454-68	0	0	13354	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	120	287
454-69	0	0	0	6566	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	146
454-70	0	0	0	0	9565	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135	197
454-71	0	0	0	0	0	9465	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	209
454-72	0	0	0	0	0	0	12416	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	213
454-73	0	0	0	0	0	0	0	7335	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	30	122
454-74	0	0	0	0	0	0	0	0	13993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	165
454-75	0	0	0	0	0	0	0	1	1	15202	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	396
454-76	0	0	0	0	0	0	0	0	1	0	10360	0	0	0	0	0	0	0	0	0	0	0	0	0	0	123	180
454-77	0	0	0	0	0	0	0	0	0	0	0	9682	0	0	0	0	0	0	0	0	0	0	0	0	0	97	145
454-78	0	1	0	0	0	0	0	0	0	0	0	0	12182	0	0	0	0	0	0	0	0	0	0	0	0	92	121
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	10443	0	0	0	0	0	0	0	0	0	0	0	114	144
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10118	0	0	0	0	1	0	0	0	0	0	186	205
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6855	0	0	0	0	0	0	0	0	0	99	118
454-82	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	11785	0	0	0	0	0	0	0	0	105	231
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12233	0	0	0	0	0	0	0	152	138
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14143	0	0	0	0	0	0	248	189
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7683	0	0	0	0	0	74	101
454-86	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8939	0	0	0	0	72	198
454-87	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10537	0	0	0	223	165
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10477	0	0	87	147
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8223	1	85	138
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5070	59	78
(close)	210	16	36	17	96	65	118	70	96	91	33	110	31	31	127	96	112	152	179	134	74	132	140	102	66		
(no match)	148	143	201	126	129	85	175	93	197	178	179	108	292	312	129	141	209	104	170	83	157	117	138	107	73		

C. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (2 nt)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13762	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	346	213
454-67	0	7494	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	143
454-68	0	0	13318	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	137	287
454-69	0	0	0	6541	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	132
454-70	0	0	0	0	9534	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	215	129
454-71	0	0	0	0	0	9438	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	116	208
454-72	0	0	0	0	0	0	12376	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	134	155
454-73	0	0	0	0	0	0	0	7318	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	40	119
454-74	0	0	0	0	0	0	0	0	13963	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	125	115
454-75	0	0	0	0	0	0	0	1	1	15159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	203	245
454-76	0	0	0	0	0	0	0	0	1	0	10338	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	161
454-77	0	0	0	0	0	0	0	0	0	0	0	9650	0	0	0	0	0	0	0	0	0	0	0	0	0	116	143
454-78	0	0	0	0	0	0	0	0	0	0	0	0	12133	0	0	0	0	0	0	0	0	0	0	0	0	124	109
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	10411	0	0	0	0	0	0	0	0	0	0	0	140	134
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10085	0	0	0	0	0	0	0	0	0	0	207	202
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6832	0	0	0	0	0	0	0	0	0	113	117
454-82	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	11736	0	0	0	0	0	0	0	0	192	162
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12187	0	0	0	0	0	0	0	175	138
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14110	0	0	0	0	0	0	260	189
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7674	0	0	0	0	0	74	105
454-86	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8906	0	0	0	0	127	155
454-87	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10494	0	0	233	169
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10444	0	107	139
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8195	1	92	141
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5052	68	79
(close)	242	86	114	68	117	67	153	72	120	142	101	116	151	166	131	94	146	150	183	137	91	145	160	113	65		
(no match)	133	86	143	91	127	94	167	102	194	151	125	118	202	193	141	153	206	129	187	85	161	133	140	114	82		

D. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (3 nt)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13566	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	639	85
454-67	0	7459	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	221	33
454-68	0	0	13260	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	314	139
454-69	0	0	0	6518	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	115
454-70	0	0	0	0	9484	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	350	23
454-71	0	0	0	0	0	9397	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	297	55
454-72	0	0	0	0	0	0	12302	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	264	54
454-73	0	0	0	0	0	0	0	7273	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	145	32
454-74	0	0	0	0	0	0	0	0	13923	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	211	51
454-75	0	0	0	0	0	0	0	0	1	15060	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	418	97
454-76	0	0	0	0	0	0	0	0	1	0	10266	0	0	0	0	0	0	0	0	0	0	0	0	0	0	308	50
454-77	0	0	0	0	0	0	0	0	0	0	0	9602	0	0	0	0	0	0	0	0	0	0	0	0	0	241	49
454-78	0	0	0	0	0	0	0	0	0	0	0	0	12082	0	0	0	0	0	0	0	0	0	0	0	0	226	34
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	10349	0	0	0	0	0	0	0	0	0	0	0	269	48
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10019	0	0	0	0	0	0	0	0	0	0	294	153
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6799	0	0	0	0	0	0	0	0	0	200	42
454-82	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	11674	0	0	0	0	0	0	0	0	357	28
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12141	0	0	0	0	0	0	0	292	43
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14049	0	0	0	0	0	0	405	77
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7632	0	0	0	0	0	155	44
454-86	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8838	0	0	0	0	250	77
454-87	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10431	0	0	0	390	57
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10395	0	0	202	65
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8148	1	201	56
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5028	136	27
(close)	369	167	210	122	237	154	317	172	268	280	209	215	299	319	240	221	325	268	346	218	198	269	258	191	126		
(no match)	37	26	76	48	28	20	48	29	64	46	42	36	78	59	60	47	58	35	52	26	77	27	70	59	29		

E. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (4 nt)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13508	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	660	90
454-67	0	7427	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	233	35
454-68	0	0	13198	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	344	141
454-69	0	0	0	6487	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	124	116
454-70	0	0	0	0	9445	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	368	21
454-71	0	0	0	0	0	9353	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	316	58
454-72	0	0	0	0	0	0	12230	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	312	38
454-73	0	0	0	0	0	0	0	7245	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	166	28
454-74	0	0	0	0	0	0	0	0	13869	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	239	44
454-75	0	0	0	0	0	0	0	0	1	14983	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	453	98
454-76	0	0	0	0	0	0	0	0	1	0	10209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	331	51
454-77	0	0	0	0	0	0	0	0	0	0	0	9572	0	0	0	0	0	0	0	0	0	0	0	0	0	257	50
454-78	0	0	0	0	0	0	0	0	0	0	0	0	12000	0	0	0	0	0	0	0	0	0	0	0	0	265	36
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	10293	0	0	0	0	0	0	0	0	0	0	0	292	50
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9963	0	0	0	0	0	0	0	0	0	0	307	172
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6765	0	0	0	0	0	0	0	0	0	210	44
454-82	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	11625	0	0	0	0	0	0	0	0	375	29
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12081	0	0	0	0	0	0	0	310	45
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13976	0	0	0	0	0	0	440	77
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7602	0	0	0	0	0	169	43
454-86	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8797	0	0	0	0	268	80
454-87	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10378	0	0	0	416	58
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10346	0	0	228	63
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8126	1	208	55
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5013	150	25
(close)	401	185	235	128	262	178	359	183	300	323	236	226	341	352	263	243	356	307	383	234	215	296	282	205	129		
(no match)	37	26	81	56	26	18	46	29	65	44	48	38	77	57	61	47	57	36	53	27	80	26	71	61	29		

F. SEARCH MOTIF = Barcode (10 nt) + Cloning-Linker (complete sequence)

	454-66	454-67	454-68	454-69	454-70	454-71	454-72	454-73	454-74	454-75	454-76	454-77	454-78	454-79	454-80	454-81	454-82	454-83	454-84	454-85	454-86	454-87	454-88	454-89	454-90	(close)	(no match)
454-66	13043	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	862	84
454-67	0	6511	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	917	61
454-68	0	0	11723	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1608	17
454-69	0	0	0	5658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	889	13
454-70	0	0	0	0	8502	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1080	16
454-71	0	0	0	0	0	8229	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1142	42
454-72	0	0	0	0	0	0	11343	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	886	76
454-73	0	0	0	0	0	0	0	6567	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	638	9
454-74	0	0	0	0	0	0	0	0	12443	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1250	52
454-75	0	0	0	0	0	0	0	0	1	11002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4102	69
454-76	0	0	0	0	0	0	0	0	1	0	9441	0	0	0	0	0	0	0	0	0	0	0	0	0	0	872	64
454-77	0	0	0	0	0	0	0	0	0	0	0	8802	0	0	0	0	0	0	0	0	0	0	0	0	0	841	31
454-78	0	0	0	0	0	0	0	0	0	0	0	0	10575	0	0	0	0	0	0	0	0	0	0	0	0	1406	40
454-79	0	0	0	0	0	0	0	0	0	0	0	0	0	9404	0	0	0	0	0	0	0	0	0	0	0	953	16
454-80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9178	0	0	0	0	0	0	0	0	0	0	957	87
454-81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6136	0	0	0	0	0	0	0	0	0	632	13
454-82	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	10651	0	0	0	0	0	0	0	0	1056	25
454-83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11062	0	0	0	0	0	0	0	1071	15
454-84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12618	0	0	0	0	0	0	1448	50
454-85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6713	0	0	0	0	0	865	25
454-86	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6070	0	0	0	0	2601	71
454-87	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9226	0	0	0	1189	65
454-88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9210	0	0	1059	37
454-89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7425	1	677	73
454-90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4509	533	10
(close)	681	395	634	331	508	476	654	422	737	718	471	454	642	644	505	488	673	608	739	436	642	647	623	422	278		
(no match)	26	22	17	21	16	34	26	15	35	10	27	15	56	27	39	40	37	23	74	36	56	47	61	58	16		