Molecular indexing for improved RNA-Seq analysis

Authors:
Bioo Scientific Corporation. 3913 Todd Lane, Suite 312, Austin, TX 78744, USA
Cellular Research, Inc. 3183 Porter Drive, Palo Alto, CA 94304, USA

Abstract

In this application note, Bioo Scientific Corporation and Cellular Research introduce a new kit for high precision gene expression analysis by RNA-Seq. We demonstrate that the NEXTflex™ qRNA-Seq™ Kit can be directly substituted into paired-end library preparation without affecting conventional RNA-Seq analysis methods. This new kit efficiently generates libraries equivalent to standard RNA-Seq libraries with sample barcodes, but with the added feature of molecular indexing. The value of using the added molecular indexing feature depends upon experimental design, data objectives, and sequencing depth for a given library complexity.

Introduction

Most modern methods for nucleic acid analysis require the use of enzyme processing, such as DNA polymerase reactions, in the sample preparation or measurement steps. For example, in microarray experiments for gene expression, the RNA samples must first be converted to cDNA and amplified to a sufficient concentration for hybridization. Likewise, for RNA-Seq experiments, a highly amplified library of DNA fragments adapted with priming sites needs to be generated[1]. Although necessary, these enzymatic steps introduce errors in the form of incorrect sequence and misrepresented copy number.

Cellular Research has developed a Molecular Indexing™ technology that labels individual DNA molecules. Individual DNA molecules of identical sequence become distinct through labeling, and can be tracked in subsequent analysis. Counting of indexed molecules in lieu of counting sequencing reads provides an absolute, digital measurement of gene expression levels, irrespective of any amplification distortions. Details of the principle of stochastic labeling and applications to identify and count individual DNA molecules have been described previously[2,3].

Conventional RNA sequencing library construction involves the ligation of a population of cDNA molecules with adaptors prior to amplification and sequencing. Any two molecules of identical sequence are indistinguishable throughout the assay. With Bioo Scientific's new NEXTflex™ qRNA-Seq™ Kit, each molecule is tagged with a molecular index randomly chosen from ~10,000 combinations so that any two identical molecules become distinguishable (with odds of 10,000/1), and can be independently evaluated in later data analysis. The kit adds no additional steps to the workflow, costs no more than a conventional library preparation kit and increases the precision of downstream analysis.

At low sequencing depths, analysis using the NEXTflex qRNA-Seq Kit is identical to conventional analysis and generates equivalent RPKM values in all applications. As sequencing depth increases, individual molecular resolution also increases. In quantitative RNA-Seq experiments, the molecular indices distinguish re-sampling of the same molecule from sampling of a different molecule[2-6]. At high sequencing depths, each molecule can be distinguished and the entire library can be analyzed to provide absolute numbers of each molecule. Resolving individual clones of molecules can also be especially useful for increasing sequencing accuracy or when identifying mutations in complex mixtures[7-9].

Molecular indexing is also particularly well suited for the analysis of small samples of limited quantity, such as in gene expression measurements within single cells. Very often, precious samples are irreplaceable and current techniques (eg. NGS, qPCR or digital PCR) are often destructive in that they consume the sample, prohibiting multiple measurements. In contrast, once molecules are labeled, the molecular indexing information is hard-coded within the population of molecules. The sample is effectively immortalized and can be re-used by amplification.

Materials and Method:

RNA Samples

Total RNA was isolated from 293F cells using the BiooPure™ RNA Isolation Reagent (Bioo Scientific). The quality of the total RNA was measured on the Agilent 2100 Bioanalyzer. Polyadenylated mRNA was purified from total RNA using the PolyA Pure Kit (Ambion) and quantitated by spectrophotometry.

Molecular indexing and qRNA-seq Library Construction

20 ng of purified mRNA was used for the construction of each paired-end library using the NEXTflex qRNA-Seq Kit (Bioo Scientific). External RNA Controls Consortium (ERCC) synthetic RNAs (Life Technologies) were spiked into the input RNA sample to serve as an internal control[10, 11]. Library preparation was performed following instructions provided in the user manual. The cDNA ligation adaptors included in the kit consist of an equimolar pool of 96 molecular indices[12]. Also included for multiplexing up to 96 libraries are sample-specific reverse PCR primers, each containing an 8 nucleotide sample barcode[13].

Library Quantitation, Quality Assessment and Sequencing

Quantitation of each library was determined by triplicate measurements with qPCR using a library quantification kit for Illumina platforms. Library quality was assessed using high sensitivity DNA chips on the Agilent 2100 Bioanalyzer. The average insert size of qRNA-Seq libraries were between 170 bp and 190 bp. Sample multiplexing was achieved by normalizing all 96 libraries to 7 nM before pooling and paired-end 2 x 150 bp sequencing on the Illumina MiSeq instrument. The molecular indexing technique is independent of sequencing platforms and may be adapted to other sequencing instruments like the Ion Torrent PGM or Proton instruments.

Sequence analysis

Adapter sequences were removed and reads were mapped using BWA to the hg19 Refseq RNA sequences downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). The 92 ERCC artificial RNA sequences were obtained from the manufacturer. A bash script was used to filter reads to require a minimum mapping quality of 30. For each transcript, the number of reads, label pairs and the mapped genome start and stop coordinates were determined. Plotting and summarization was performed in R.

Results and Discussion

A summary of the library construction steps is depicted in Figure 1. The method is identical to typical RNA-Seq library preparation protocols except for the inclusion of a set of 96 distinct molecular indices on the sequencing adaptors. Use of randomized bases as molecular barcodes has been reported[4, 5, 7-9], but some sequences can be problematic in PCR, and sequencing errors in the random barcodes can cause ambiguity in data analysis. We apply labels consisting of a distinct set of error-correcting 8-nucleotide barcode tags[12]. In the ligation reaction, these 96 adaptors are present in vast molar excess over the concentration of the cDNA fragments, and therefore serve as a non-depleting reservoir of molecular labels. Each end of a cDNA is ligated to a single label from this pool of 96 adaptors at random. The paired-end molecular indexing strategy greatly reduces the number of distinct adaptors required, yet allows for as many as 9,216 possible combinations (96 x 96) across both ends. Paired-end reads reveal the chosen label on each end along with adjoining cDNA sequence, and reads from PCR duplicated clones are readily distinguished (Figure 2). In addition to indexing DNA fragments at the molecular level, we also applied sample-specific barcodes during the library preparation PCR step.

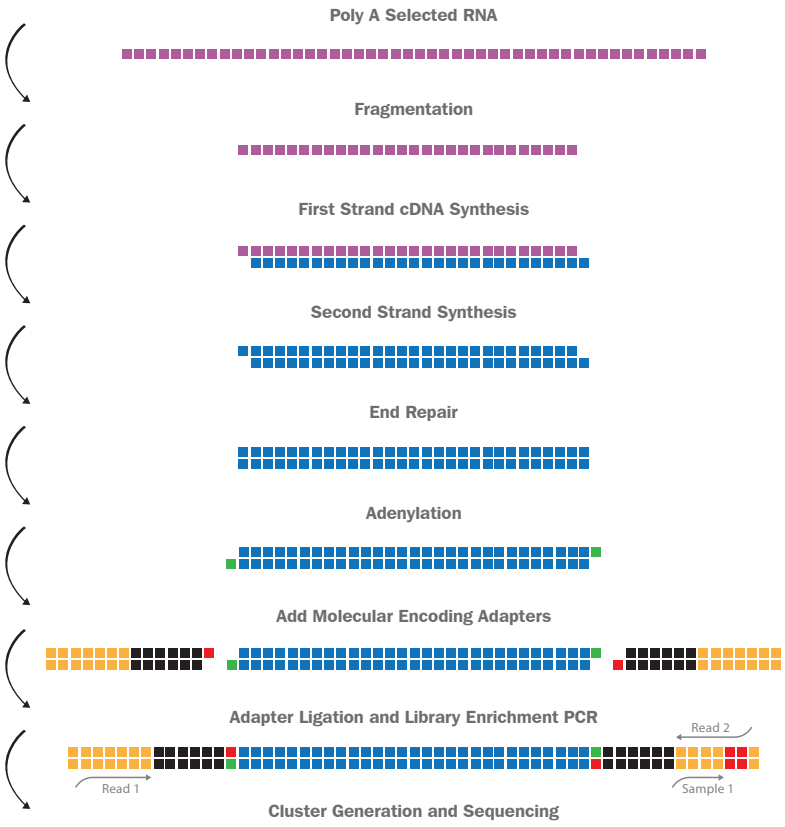

Figure 1. Outline of the library construction workflow. Isolated mRNAs (purple bars) are fragmented and cDNA (blue bars) is generated by reverse transcription with random primers followed by second strand synthesis. After purification, the cDNA ends are repaired, and an A overhang is added to enable adaptor ligation. On each end of a cDNA fragment, one of 96 possible adaptors (black bars) is ligated. Unreacted excess adaptors are purified away and the resulting cDNAs are enriched by PCR using common primer sequences present on the sequencing adaptors (yellow bars). If desired, sample barcodes can be added to the PCR primers (red bars).

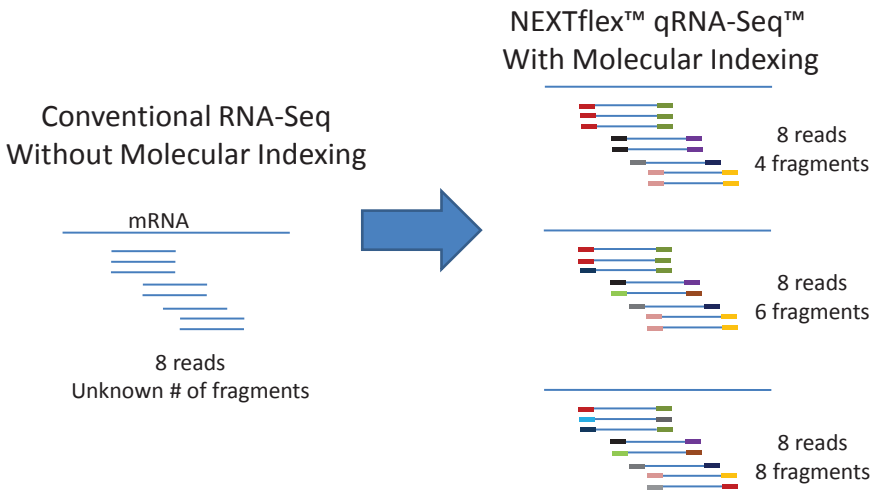

Figure 2. An illustration with 8 mapped reads to an mRNA transcript either with or without molecular indexing. Individual molecular indices at fragment ends are represented by different colors. The number of cDNA fragments represented by the reads is unknown without molecular indexing, but can be determined with molecular indexing.

To evaluate the sequenced libraries for their ability to measure RNA abundance accurately, we determined the RPKM values for the set of ERCC spike-in RNAs. These control RNAs consist of 92 poly-adenylated artificial transcripts spanning a 10^6-fold concentration range. Observed RPKMs correlate well with spike-in concentration in general. However, sampling errors increase at very low RNA concentrations as expected (Figure 3).

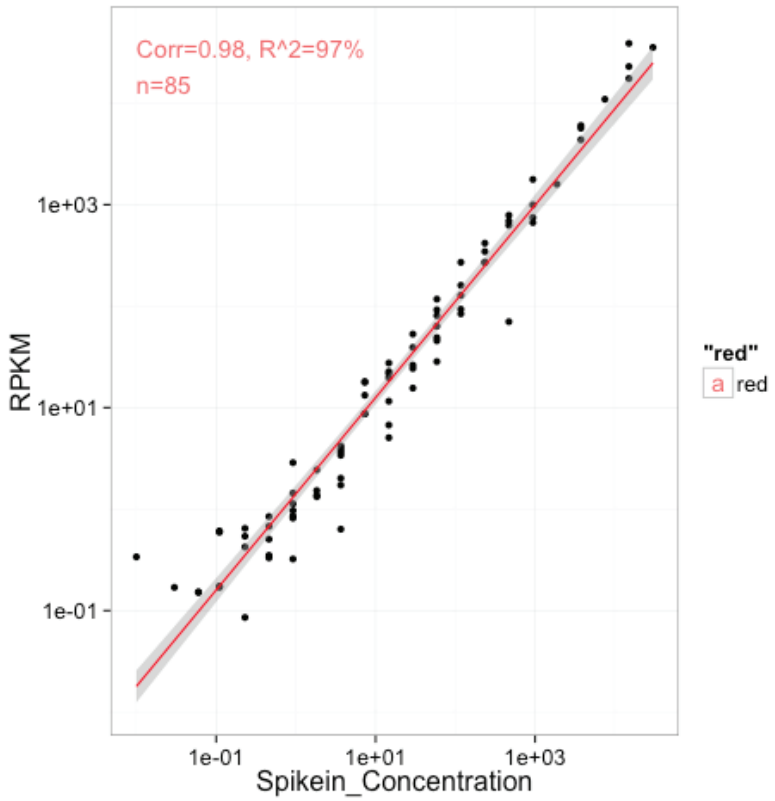

Figure 3. A plot of RPKM values (Y-axis) for 85 detected ERCC spike-in Control RNAs. Relative concentrations supplied by the manufacturer are shown on the X-axis.

In conventional RNA-seq analysis, RPKM values are used to determine the relative concentrations of each gene transcript in the sample[1]. An inherent weakness of RPKM analysis is that cDNA fragments that amplify more efficiently will unavoidably result in a higher number of reads than cDNAs that do not amplify as well during the library construction PCR step. Therefore, when multiple reads mapping to the same transcript are encountered, it is not possible to determine whether sequenced reads originate from the same or different cDNA molecule. As a remedy to this re-sampling problem, many researchers evaluate whether or not each read has the same start and stop mapping coordinates. Reads with identical start and stop positions are usually assumed to be clonal duplicates derived from the same parent molecule. However, this assumption is not always true because start and stop analysis does not reveal the extent of bias in the RNA fragmentation, random priming or reverse transcription processes that may lead to similar cDNA ends for different copies of RNA molecules. Furthermore, theoretically available start and stop combinations decrease with smaller RNA size and larger (or more narrowly selected) cDNA fragment lengths. For example, start and stop analysis has limited utility for microRNA sequencing as the short transcripts predominantly result in identical read ends. cDNA libraries prepared from small input sample amounts requiring greater levels of PCR amplification are more prone to re-sampling biases, and the number of apparently duplicated reads also increase for a more abundant gene or with deeper sequencing.

As a solution for measuring errors that arise from re-sampling, the NEXTflex qRNA-Seq Kit labels each cDNA fragment end with a randomly chosen molecular index before any amplification steps. An archive of Digitally Encoded DNA™ molecules is created, and can be amplified as desired. The molecular indices used in the NEXTflex qRNA-Seq consist of a set of 96 sequence-labeled ligation adaptors (Figure 1). For a random stochastic labeling process, the number of label pairs chosen (k) can be calculated from the number of cDNA molecules labeled (n) using the equation $k = m(1- e^{-(n/m)})$. A plot of k versus n is shown in Figure 4 for a set of 9,216 total labels (m). This set of labels provides sufficient molecular indexing capacity for a majority of RNA-Seq experiments where the number of reads belonging to any given gene fragment is very unlikely to exceed a few thousand. When a library with molecular indexing is sampled deeply, absolute numbers of molecules are revealed. A more comprehensive report of the use of molecular indexing for RNA-Seq is in preparation[14]. However, even in lightly sampled libraries (as shown here,) valuable new data insights are obtained.

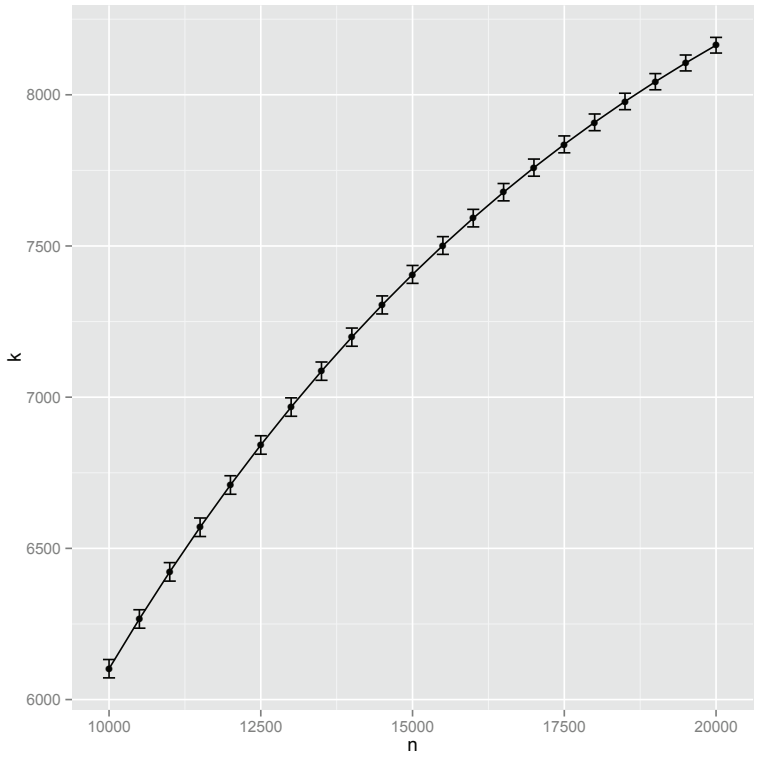

Figure 4. Average number of distinct paired-end labels (k) expected for a given number of target cDNA molecules (n) calculated using a total of 9,216 possible label pairs (m). Error bars represent one standard deviation away from the calculated mean values.

At low sequence sampling depths, reads are mostly derived from distinct molecules and counting reads or counting molecules is largely equivalent. As the sampling depth increases, so does the likelihood of observing multiple reads amplified from the same molecule (clonal duplicates). Although it is common practice to group reads with the same start and stop sites and assume they are identical molecules, one is unable to verify that they are indeed clonal duplicates. Here, we show the differences in counts using either conventional analysis or molecular indexing (figure 5). As expected, with higher read depth, start and stop analysis becomes increasingly limited in the ability to count different molecules. On the other hand, molecular indexing enables a significantly more reliable level of informed analysis, clearly identifying the true extent of re-sampling.

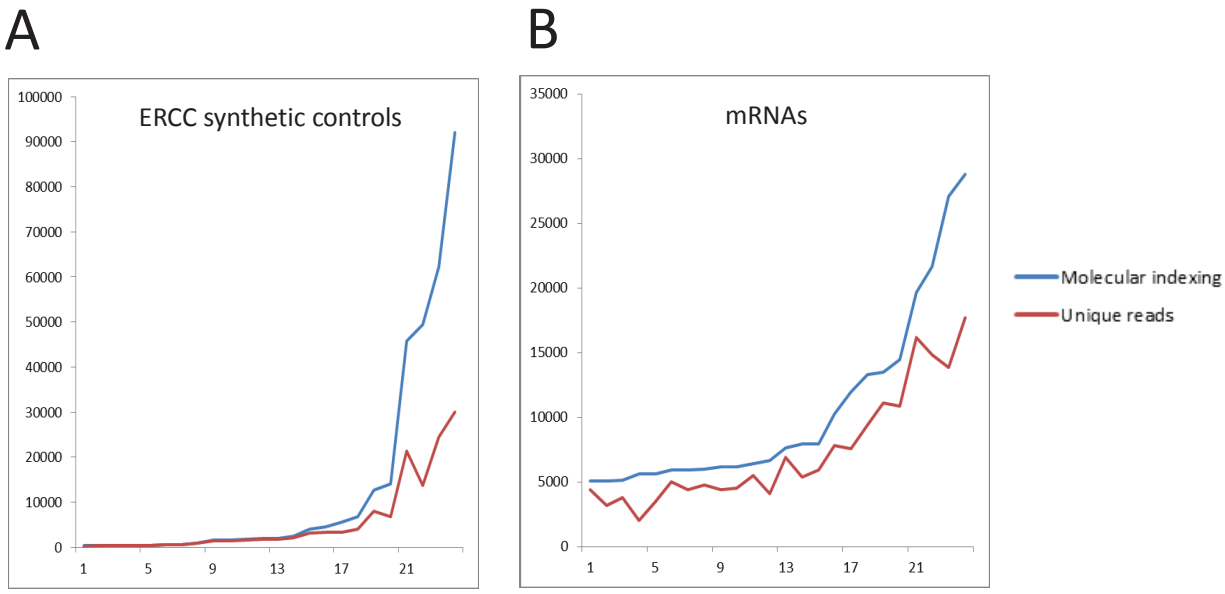

Figure 5. The number of unique fragments as determined by reads with unique start and stop sites (red line) or by molecular indexing (blue line). Data shown is for a set of24 ERCC controls (A), or 24 mRNAs (B).

The NEXTflex qRNA-Seq Kit generates a library equivalent to a conventional paired-end RNA-Seq library – in terms of cost and workflow - yet provides additional molecular information allowing for the determination of the degree of re-sampling, the ability to detect and distinguish sequencing errors, the identification of mutations in mixtures, and with sufficient depth, absolute counting of each molecule in the library.

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8. PubMed PMID: 18516045.

2. Fu GK, Hu J, Wang PH, Fodor SP. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A. 2011 May 31;108(22):9026-31. PubMed PMID: 21562209; PubMed Central PMCID: PMC3107322.

3. Fodor SP, Fu GK; Digital Counting of Individual Molecules by Stochastic Attachment of Diverse Labels. USA US 20110160078 A1. 2011.

4. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 2011 Jul;39(12):e81. PubMed PMID: 21490082; PubMed Central PMCID: PMC3130290.

5. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2011 Nov 20;9(1):72-4. PubMed PMID: 22101854.

6. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. 2012 Jan 24;109(4):1347-52. PubMed PMID: 22232676; PubMed Central PMCID: PMC3268301.

7. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011 Jun 7;108(23):9530-5. PubMed PMID: 21586637; PubMed Central PMCID: PMC3111315.

8. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. 2011 Dec 13;108(50):20166-71. PubMed PMID: 22135472; PubMed Central PMCID: PMC3250168.

9. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012 Sep 4;109(36):14508-13. PubMed PMID: 22853953; PubMed Central PMCID: PMC3437896.

10. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Warrington JA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R, External RNA Controls Consortium. The External RNA Controls Consortium: a progress report. Nat Methods. 2005 Oct;2(10):731-4. PubMed PMID: 16179916.

11. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics. 2005 Nov 2;6:150. PubMed PMID: 16266432; PubMed Central PMCID: PMC1325234.

12. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. 2008 Mar;5(3):235-7. PubMed PMID: 18264105; PubMed Central PMCID: PMC3439997.

13. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010 Feb;7(2):111-8. PubMed PMID: 20111037.

14. Fu G, Wilhelmy J, Xu W, Xiao WZ, Mindrinos M, Davis R and Fodor SPA. Manuscript in preparation.

Bioo Scientific Corporation
3913 Todd Lane, Suite 312
Austin, Texas 78744

www.biooscientific.com
info@biooscientific.com
1.512.707.8993