Directional qRNA-Seq: Combining the Power of Stranded RNA-Seq with the Quantitative Precision of Molecular Labels

Authors: Radmila Hrdličková[a], Jiří Nehyba[a], Weihong Xu[b]

[a] Bioo Scientific Corporation, 7050 Burleson Rd., Austin, TX 78744, USA

[b] Stanford Genome Technology Center, 3165 Porter Dr., Palo Alto, CA 94304

ABSTRACT

Bioo Scientific introduces a novel product for advanced RNA-Seq library preparation that features strand-specific RNA sequencing and correction of PCR amplification bias by molecular indexing: the NEXTflex™ Rapid Directional qRNA-Seq™ Kit. We demonstrate the utility of this product for identification of novel non-coding RNAs and unbiased measurement of RNA expression. The NEXTflex Rapid Directional qRNA-Seq Kit, available together with new software for analysis of molecular indexed data, provides a convenient all-in-one solution for accurate analysis of gene expression.

INTRODUCTION

RNA-Seq is a powerful tool for transcriptome analysis. However, data generated from standard RNA-Seq protocols lack two important pieces of information about RNA expression—the DNA strand from which RNA transcripts were derived and the precise abundance of those transcripts.

While the genomes of thousands of species have already been sequenced, the assembly and annotation of most transcriptomes are still incomplete (1-3). The major focus of transcriptome analyses is on protein-coding transcripts, which represent less than 2% of the genome. These genes are transcribed from either of two strands of DNA, but their sequence can often be predicted based on open-reading-frame and splicing analysis even in the absence of strand information (4). However, in recent years, several studies have shown that a significant portion of the genome (more than 80%) is transcribed (5, 6). Some of these transcripts are non-coding RNA (ncRNAs), which often play major regulatory roles at the transcriptional and/or translational level (7). While ncRNAs are readily detected by Next Generation Sequencing (NGS), determining from which DNA strand they were derived can be difficult, especially when these transcripts are not spliced (8, 9). New RNA-Seq techniques allow for the preservation of strand-specific information during library construction (10). One common and effective solution, incorporated into the NEXTflex Rapid Directional RNA-Seq Kit, involves a modification of second strand synthesis by incorporation of dUTPs, which is followed by degradation of this strand by uracil-glycosylase prior to PCR amplification (11). As a result, only the first cDNA strand is sequenced, and strandedness is maintained by directional adapter ligation, i.e. such that the Illumina P7 adapter is ligated to the 3' end of the cDNA strands and the P5 adapter is ligated to the 5' end.

Gene expression value is proportional to the number of sequencing reads mapped to the gene relative to the whole transcriptome. However, PCR amplification introduces bias because some fragments are preferentially amplified, so the frequencies of the transcripts in the original RNA sample are not faithfully reproduced. To correct this bias, individual fragments can be randomly tagged either during reverse transcription or through adapter ligation (12-20), a strategy referred to as molecular indexing. Bioo Scientific, in collaboration with Cellular Research, offers the only commercial kit to use molecular labels to provide an unbiased quantification of RNA expression in RNA-Seq libraries: the NEXTflex qRNA-Seq Kit.

Here we describe a new kit, which provides both directionality and molecular indexing in a single protocol: the NEXTflex Rapid Directional qRNA-Seq Kit.

METHODS

1. Library Construction

Three different types of libraries (Table 1) were constructed using 20 ng of poly(A)+ mRNA (BioChain Institute, Newark, CA) from the MCF7 breast cancer cell line. Each library was evaluated by Agilent Bioanalyzer for quality and by qPCR for quantity.

2. Molecular Labels

The NEXTflex Rapid Directional qRNA-Seq Kits contain a set of 96 distinct molecular labels incorporated into new adapter sequences (NEXTflex™ Molecular Index Adapters). Each label consists of an 8 nucleotide barcode tag. During the ligation reaction, each cDNA fragment randomly ligates to two labels from this pool, one on each end, resulting in a total of 96 x 96 = 9,216 possible combinations across both ends. For every clone sequenced, paired-end reads reveal the label on each end along with the adjoining cDNA sequence. This makes molecules of identical sequence distinguishable. In our experiments we examined the dynamic range of the NEXTflex Molecular Index Adapters (dqy and qy) (Fig. 1). To determine read distribution, FASTQ files with sequence reads were trimmed to nine 5' nucleotides, converted to FASTA using Galaxy tools, and the labels were counted by custom designed Bourne-again shell script.

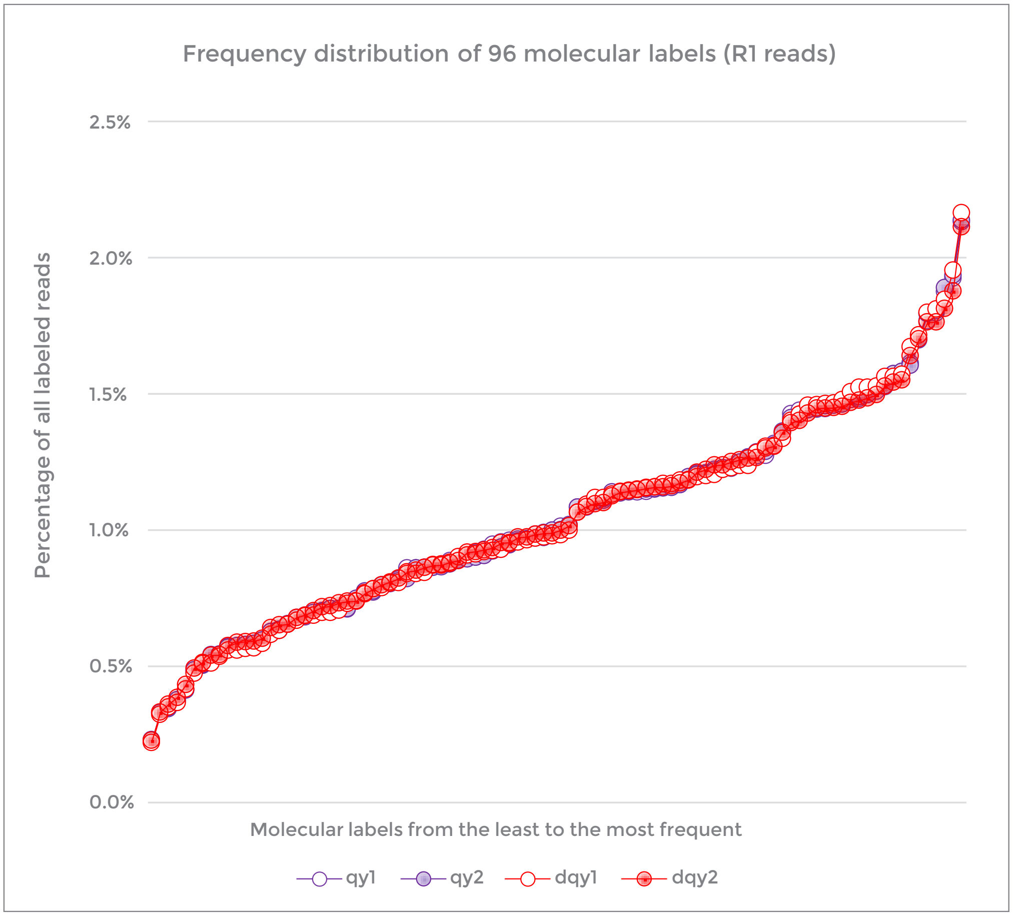

Figure 1. Read frequency of individual NEXTflex Molecular Index Adapters (dqy, qy).
Libraries: the NEXTflex Rapid Directional qRNA-Seq Kit (dqy) and the qRNA-Seq protocol with NEXTflex Molecular Index Adapters (qy). Each protocol has two duplicate libraries indicated by the same color. The labels were sorted from the least to the most frequent, separately for each library. The R2 labels had a similar distribution as R1 (data not shown).

3. Sequence Analysis

Paired-end 2×100 bp sequencing was performed on Illumina HiSeq 2500 instruments at the University of Texas (Austin, TX) and Texas A&M University (College Station, TX) core sequencing facilities, where reads were de-multiplexed. All FASTQ files contained only the expected barcoded reads. Read quality was further verified by FastQC (Babraham Bioinformatics, UK). Seven to eight low quality 3' nucleotides were trimmed from all sequences (Trimming tool based on FASTX-tool kit, http://usegalaxy.org). Nine 5' nucleotides comprised of molecular label were removed before alignment to the transcriptome or genome either by trimming tool or by alignment software. Reference hg19 transcriptome file was assembled by pooling the UCSC RefSeq hg19 transcriptome (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/refMrna.fa.gz) with all 15 mitochondrial protein-coding and mitochondrial rRNA sequences retrieved from Ensembl GRCh37.75 reference sequences. Reads were aligned by Bowtie2 (21). Genomic alignment was performed by TopHat2 to the hg19 reference genome (22). Random selection of reads was executed by a pipeline of Galaxy tools: FASTQ joiner – FASTQ to tabular – Select random lines – Tabular to FASTQ – FASTQ Groomer – FASTQ splitter. Reads in exons were counted by htseq-count script (23, 24) using Ensembl GTF annotation file Homo_sapiens.GRCh37.56.gtf. Duplicate reads were detected by the publicly available dqRNASeq script.

RESULTS AND DISCUSSION

The NEXTflex Rapid Directional qRNA-Seq Kit produces both stranded and quantitative data. Three different libraries were constructed (Table 1) to assess the effectiveness of the NEXTflex stranded RNA-Seq kits in retaining directionality and correcting PCR bias. All three protocols were evaluated for depth, distribution of reads and complexity.

Table 1. RNA-Seq libraries constructed for analysis

Kit Name	Cat. Number	Name	Dir.[a]	MI[b]	Y[c]
NEXTflex™ Rapid Directional qRNA-Seq™ Kit	5130-01D(02D-05D)	dqy	+	+	+
NEXTflex™ Rapid Directional RNA-Seq Kit	5129-07(08)	dy	+	-	+
Non-directional RNA-Seq protocol		qy	-	+	+

[a] *Directional*
[b] *Molecular indexing*
[c] *Y-shaped adapters*

1. NEXTflex Rapid Directional qRNA-Seq Kit (dqy) provides accurate fragment-level detection of duplicates. To correct for PCR bias, the identification of unique start and stop coordinates (USS) is used to eliminate duplicates. This method works based on the assumption that the initial chemical or mechanical fragmentation of RNA is random and therefore all fragments generated in this way would have largely different start and/or stop sites. Consequently, reads with identical start/stop sites are collapsed into a single read. Previous work has shown that many original fragments have identical start and stop sites (25). As a result, the application of the USS method alone incorrectly eliminates distinct original fragments.

The NEXTflex Rapid Directional qRNA-Seq Kit (dqy) randomly ligates 96 molecular labels (also known as stochastic labels, or STLs) at both ends of each cDNA fragment prior to PCR amplification, allowing for 9,216 unique combinations of labels. The chance of two fragments having both the same start/stop sites and molecular indices is extremely small; thus molecular indexing using stochastic labels (STL) can be used in combination with the USS method to distinguish identical but distinct starting molecules from true PCR duplicates (16, 17). This retains reads that would otherwise be discarded as PCR duplicates, allowing for a more accurate analysis. Combining both stochastic label and start/stop correction is necessary because the application of molecular indices alone does not provide enough combinations to accurately distinguish duplicate reads for highly expressed genes.

For deconvolution of unique read fragments, Bioo Scientific now offers a dqRNASeq script, created by Weihong Xu of the Stanford Genome Technology Center and licensed under the GNU General Public License (GPL). Using read pairs aligned to transcripts and FASTQ files, this script will generate a table listing fragments, their start/stop sites in transcripts, and their molecular labels (STL). The script will also generate a table listing the total number of read pairs per transcript and the number of read pairs after STL, USS, and STL/USS correction.

Using the dqRNASeq script, we demonstrated that a substantial portion of RNA-Seq read pairs aligned to the transcriptome were improperly removed based on USS alone (Fig. 2). Analysis by USS demonstrated that 20-30% of properly aligned fragments are removed based on identical start and stop sites (USS). However, additional analysis of molecular indices revealed that many of these fragments had different combinations of molecular indices (USS + STL) and therefore should have been retained, shrinking the number of duplicates to 5-10% of read pairs. The difference between USS correction and combined USS + STL correction increases with the amount of read pairs assigned to a given transcript. The percentage of fragments with the same start and stop sites retained by the application of molecular indexing depends on the starting material, the depth of sequencing, and the method of library preparation.

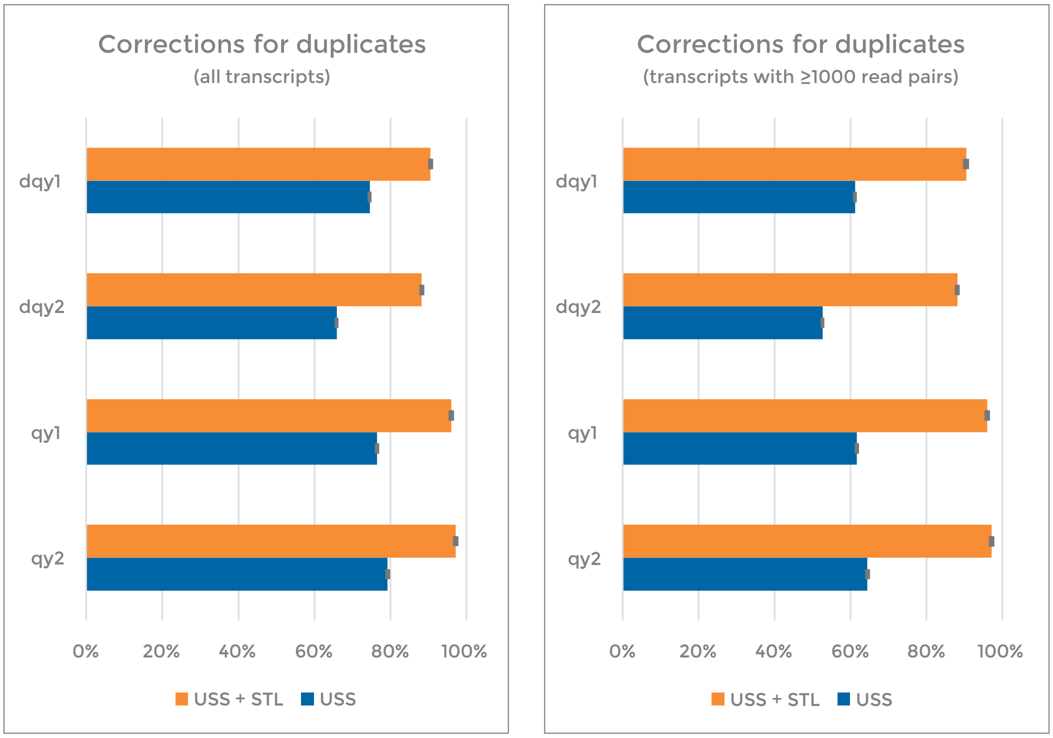

Figure 2. The number of unique fragments as determined by unique start and stop correction (USS) and a USS correction combined with molecular labels (USS + STL).
Reads were aligned to the transcriptome using Bowtie2. The number of unique fragments was determined for the read pairs with mapping quality MAPQ ≥ 30 using dqRNA script and a table of 96 molecular labels. Correction for all detected transcripts is compared with the correction for highly expressed genes (≥1,000 read pairs). Libraries: the NEXTflex Rapid Directional qRNA-Seq Kit (dqy) and a non-directional RNA-Seq Protocol (qy). Each library was tested in duplicate, and each column is an average of the same library sequenced in two different lanes (error bars indicate SD).

2. Directional and non-directional protocols deliver a similar distribution of reads. To quantify RNA sequences in libraries produced by the three different protocols, reads were aligned to the human reference genome and the distribution of reads over genomic features was calculated (Fig. 3). The percentage of reads aligning to exons (60%) and introns (25-30%) was consistent across all protocols. The number of reads aligning to introns reflects the detection of un-spliced transcripts and also expression of unannotated non-coding RNA.

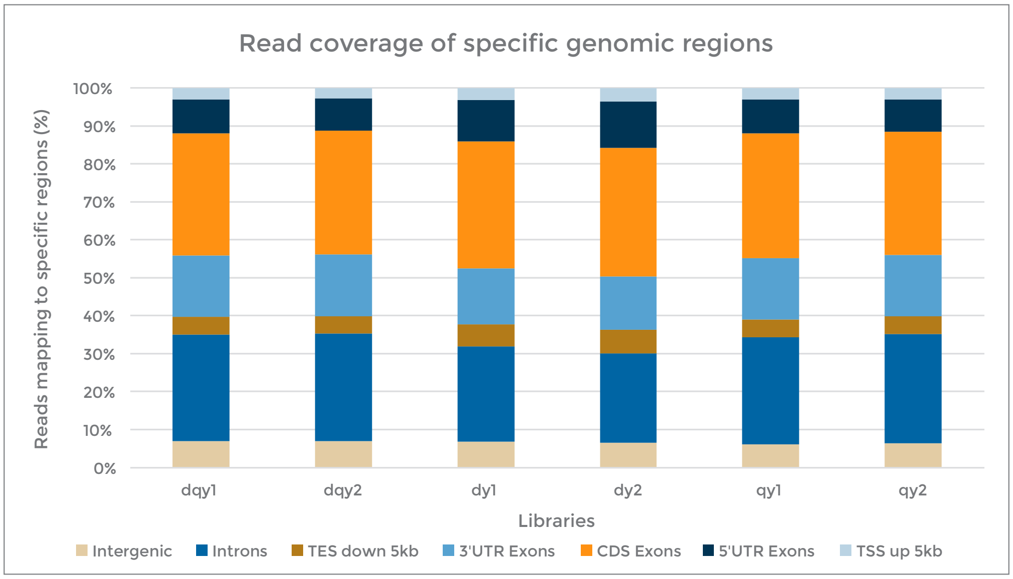

Figure 3. The distribution of sequencing reads along the genome.
Alignment to the hg19 reference genome was performed using TopHat2, and the distribution of reads over genomic features was calculated with the RseQC package. Libraries: the NEXTflex Rapid Directional qRNA-Seq Kit (dqy), *the NEXTflex Rapid Directional RNA-Seq Kit* (dy) *and the non-directional RNA-Seq protocol* (qy). *Each experimental protocol includes two duplicate libraries.*

3. The NEXTflex Rapid Directional qRNA-Seq libraries retain complexity of the original NEXTflex Rapid Directional RNA-Seq protocol. The complexity of a library—the representation of RNA transcripts related to the total number reads—is known to drop with the reduction of input material, but can also be protocol-dependent (24). To test the ability of the NEXTflex Rapid Directional qRNA-Seq Kit (dqy) to form libraries of sufficient complexity, protocols were compared and the frequency of individual transcripts determined (Fig. 4). These results confirmed that the complexity of the library was not compromised by the introduction of molecular labels. All directional libraries (dqy1, dqy2, dy1, dy2) had smaller numbers of transcripts detected, mostly due to a decrease in genes with background level of expression (detected by one or two read pairs). This is expected in directional libraries due to a reduction of available PCR templates by the destruction of the second strand. Non-directional libraries (qy1, qy2) have slightly higher complexity with about 15% more genes detected by the same number of reads.

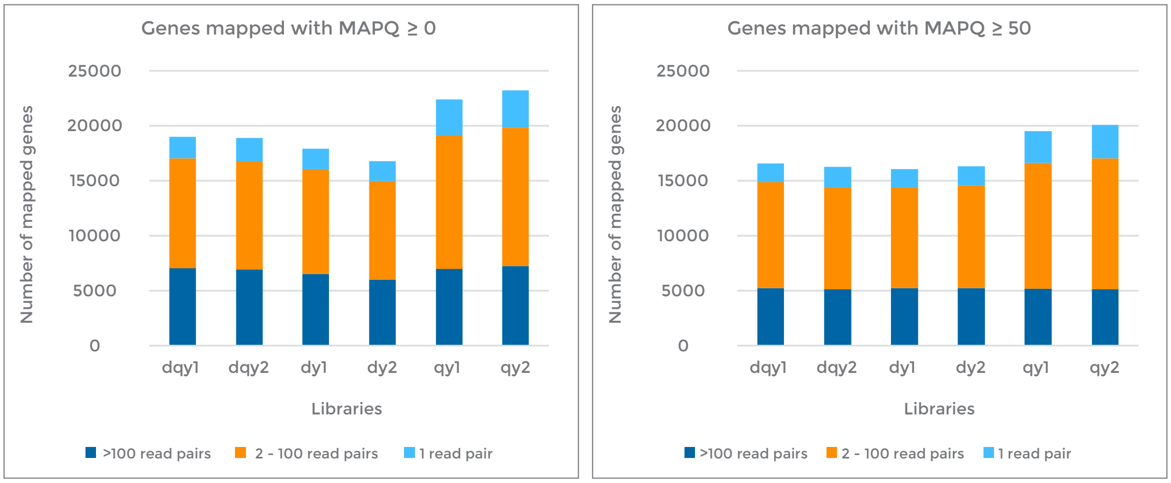

Figure 4. Complexity of libraries.
The same number of read pairs (about 7.5 million) was randomly selected from seven libraries to match the number of pairs in the smallest eighth library (dqy1). Libraries are described in Fig. 2. Reads were aligned to the genome by TopHat2 and complexity above several quality cutoffs: MAPQ ≥ 0, 1, 3, or 50 was determined. Results are shown for MAPQ ≥ 0 and 50. Numbers of mapped genes were normalized by the number of mapped reads. The y axis corresponds exactly to the number of detected genes in one of the libraries (dy2), and the gene numbers in other libraries were adjusted accordingly.

4. The NEXTflex Rapid Directional qRNA-Seq Kit retains strandedness and reveals non-coding RNAs. To demonstrate that the new directional qRNA kit retains the same strandedness delivered by the standard Bioo Scientific NEXTflex Rapid Directional RNA-Seq Kit (dy), the data produced by these two protocols were directly compared (Fig. 5). Data produced using the non-directional RNA-Seq Protocol (qy) served as negative controls. The results showed that over 98.8% of reads from both the NEXTflex Rapid Directional qRNA-Seq Kit (dqy) and the NEXTflex Rapid Directional RNA-Seq Kit (dy) mapped to the predicted DNA strand. A non-directional protocol (qy), as expected, did not discriminate between strands. The NEXTflex Rapid Directional qRNA-Seq Kit (dqy) retained even higher strandedness (over 99.8%) when considering ERCC RNA Spike-In Control Mixes (Life Technologies, Grand Island, NY) (Table 2).

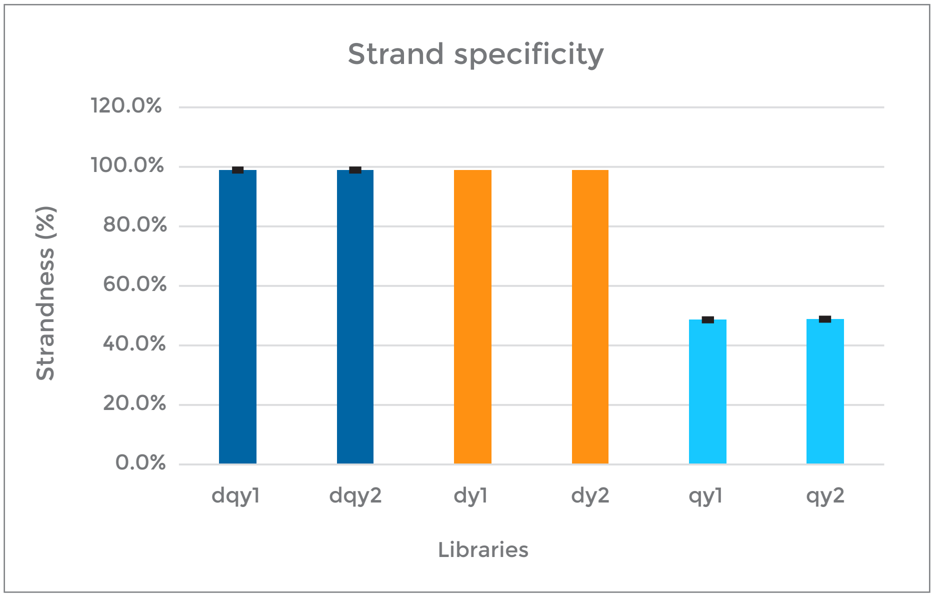

Figure 5. Determination of directionality of the sequence reads by alignment to RefSeq transcriptome.
Libraries: The NEXTflex Rapid Directional qRNA-Seq Kit (dqy), the NEXTflex Rapid Directional RNA-Seq Kit (dy) and a non-directional RNA-Seq protocol (qy). Each experiment has two duplicate libraries indicated by the same column color and each column (except dy libraries) is an average of the same library sequenced in two different lanes (error bars indicate SD). Paired-end sequences were aligned by Bowtie2 separately to the reverse strand of the RefSeq transcriptome as well as to both strands simultaneously. The alignment statistics determined by flagstat of SAMtools were compared. Results are shown as percentage of reads aligned to reverse strand (alignment to this strand reflects detection of the mRNA sense strand) compared to the simultaneous alignment to both strands.

Table 2. Strandedness of RNA spiked with ERCC RNA Spike in Control Mix in two libraries prepared by NEXTflex Rapid Directional qRNA-Seq Kit (dqy).

Library[a]	#1	#2
ERCC forward	0.2%	0.2%
ERCC reverse	>99%	>99%
HSA forward	2%	2%
HSA reverse	99%	99%

[a] Libraries were prepared from MCF7 mRNA spiked with ERCC RNA Spike-in Control Mix (Life Technologies, Grand Island, NY). Strand specificity was determined as described in Fig. 5, HSA – human mRNA transcripts.

The slight increase (1%) in the strandedness of ERCC compared to human gene transcripts suggests the presence of genuine RNA complementary to human transcripts. Accordingly, a small percentage of transcriptome-aligned read pairs (2.2–2.7%) aligned to the forward strand of human hg19 transcriptome, revealing the detection of putative mRNA antisense strands. In 60% of these read pairs, the same pair also mapped to the opposite (reverse) strand of the transcriptome, demonstrating the detection of matched sense-antisense transcripts. Therefore, the directional protocol is capable of discovering potentially functional antisense RNAs.

Candidates of novel ncRNA with regulatory functions lie in non-coding regions where reads map to promoters and introns (9). Here we present two examples of the utility of directional RNA-seq for the detection of ncRNA (Fig. 6). The first example is the non-coding RNA XR_241724 located in the intron of IRX4 gene and non-coding RNA NR_109912 located in the IRX4 promoter region (Fig. 6A-B). Both are transcribed in the opposite orientation of the IRX4 ORF. Their sequence reads can easily be distinguished from IRX4 transcript with directional sequencing (Fig. 6A) but not with non-directional approaches (Fig. 6B), especially when splicing events are not detected at the read. The second example is the CLPTM1L promoter-associated (pa) noncoding RNA (Fig. 6C-D). With non-directional sequencing (Fig. 6D) it is difficult to determine if the CLPTM1L promoter-mapped reads are derived from an antisense gene or unannotated upstream exons of CLPTM1L. The CLPTM1L pa ncRNA is not yet annotated in either the NCBI or Ensembl databases, suggesting that directional RNA-Seq can be used to discover previously unknown ncRNAs. This analysis demonstrates the power of employing directional protocols for the preparation of RNA-Seq libraries and as a tool to identify novel ncRNAs.

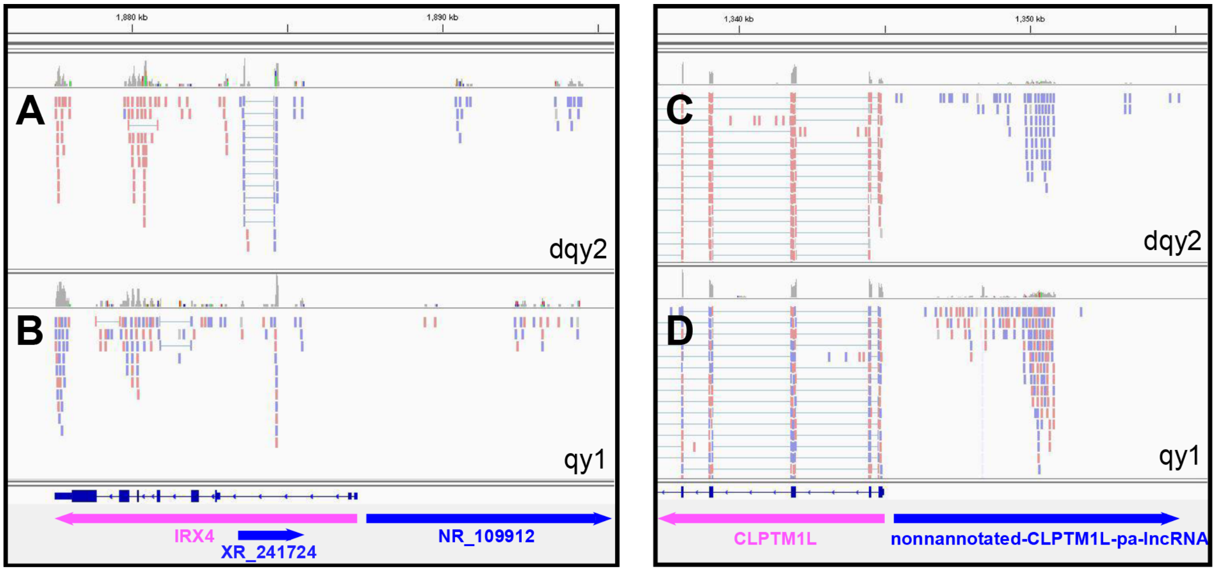

Figure 6. Noncoding RNA transcript detection with the NEXTflex Rapid Directional qRNA-Seq Kit (dqy) as opposed to a non-directional RNA-Seq protocol (qy).
Reads from directional (dqy2) and non-directional (qy1) libraries were aligned to human hg19 genome with TopHat2 and mapped reads were visualized with IGV. The reads are colored according to the direction of the first read in each pair.

CONCLUSION

The introduction of a new protocol, combining the advantages of stranded selection with unique fragment calling using molecular labels, sets a new standard in high precision biology. The NEXTflex Rapid Directional qRNA-Seq Kit provides a complete solution with both production-grade reagents and robust analytical software for studying non-coding RNA and the absolute quantification of gene expression.

References

1. Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12:671-682.
2. Florea LD, Salzberg SL. 2013. Genome-guided transcriptome assembly in the age of next-generation sequencing. IEEE/ACM Transaction on Computation Biology and Bioinformatics 10:1234-1240.
3. Góngora-Castillo E, Buell CR. 2013. Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Nat. Prod. Rep. 30:490-500.
4. Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931-945.
5. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.
6. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22:1760-1774.
7. Cech TR, Steitz JA. 2014. The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157:77-94.
8. Gelderman G, Contreras LM. 2013. Discovery of postranscriptional regulatory RNAs using next generation sequencing technologies. Method Mol. Biol. 985:269-295.
9. Ernst C, Morton CC. 2013. Identification and function of long non-coding RNA. Front. Cell. Neurosci. 7:1-9.
10. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7:709-715.
11. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 37:e123.
12. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11:163-166.
13. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl. Acad. Sci. USA 108:20166-20171.
14. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of National Academy of Sciences of the United States of America 108:9530-9535.
15. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2011. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9:72-74.
16. Fu GK, Hu J, Wang PH, Fodor SP. 2011. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl. Acad. Sci. USA 108:9026-9031.
17. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SP. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl. Acad. Sci. USA 111:1891-1896.
18. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl. Acad. Sci. USA 109:14508-14513.
19. Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl. Acad. Sci. USA 109:1347-1352.
20. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. 2011. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 39:e81.
21. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Genome Biol. 14:R36.
22. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.
23. Anders S, Pyl PT, Huber W. 2014. A Python framework to work with high-throughput sequencing data. BioRxiv doi: 10.1101/002824.
24. Bhargava V, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. 2014. Technical variations in low-input RNA-seq methodologies. Scientific Reports 4:1-10.
25. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovsky SL. 2014. Non-random DNA fragmentation in next-generation sequencing. Scientific Reports 4:1-6.