Induro™, a Novel Reverse Transcriptase for Nanopore Direct RNA Sequencing with Significantly Improved RNA 5' Coverage

Luo Sun, Yan Xu, Bradley W. Langhorst, Jian Sun, Nicole M. Nichols, Eileen T. Dimalanta, and Theodore B. Davis

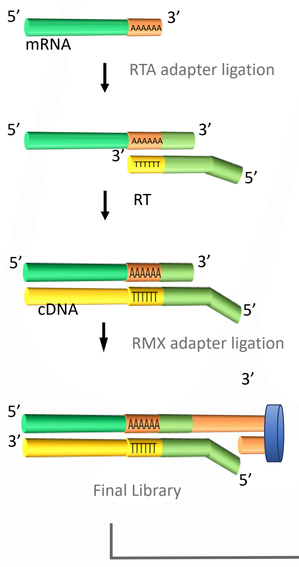
New England Biolabs, Inc.

INTRODUCTION

Transcriptome analysis based on cDNA sequencing of direct or PCR amplified cDNA has limitations, including generation of cDNA chimeras during library preparation, artifacts associated with template switching, DNA contamination and size bias during PCR amplification. Recent advances from Oxford Nanopore Technologies® (ONT) have enabled direct RNA sequencing without PCR. This method not only interrogates full-length RNA molecules and generates higher resolution data on complex isoforms, but also allows for the detection of RNA modifications. This opens up new opportunities for studying RNA stability and regulation. To perform direct RNA sequencing, one approach is to first convert RNA into an RNA/cDNA hybrid molecule using a reverse transcriptase (RT). The RNA/cDNA hybrid molecule reduces RNA secondary structure formation, leading to higher sequencing throughput and longer sequencing reads without sacrificing RNA modification information. To accomplish this, an ideal reverse transcriptase would be highly processive and generate high cDNA yields.

Here we introduce Induro, a novel reverse transcriptase that delivers superior performance in the cDNA synthesis reaction. Using this novel RT, we generated a RNA/cDNA hybrid library starting with mouse brain polyA-enriched RNA. After ligation with ONT sequencing adapter following the SQK-RNA002 protocol, this library was sequenced on the GridION. When compared with other commonly used RTs, Induro produced longer average read lengths and better 5' end coverage of RNA transcripts. Similar results were also obtained by performing direct sequencing on a cDNA library generated using this RT enzyme. In summary, Induro is a significant improvement compared to other available RT enzymes. It can produce longer cDNA molecules, and exhibits higher cDNA yields, better full-length coverage, leading to more accurate transcriptome analysis.

METHODS



Preparation time

SSIII
Total: 2 hrs
Hands on: 30min

Induro
Total: 1.2 hr
Hands on: 30min

Oxford Nanopore (ONT) RNA Library prep

RTA ligation → RT reaction → Beads clean up → RMX ligation → Beads clean up

- 500 ng input mRNA
- Ligate RTA adapter (SQK-RNA002 protocol) using NEBNext ® Quick Ligation Module (E6056)
- RT reaction: SuperScript™ III (SSIII, 50 min at 50ºC), Induro (15 min 60ºC, 10 min 70ºC)
- Ligate RMX adapter using NEBNext Quick Ligation Module (E6056)

Oxford Nanopore (ONT) multiplex RNA Library Prep

RTA barcode ligation → RT reaction → Pool Beads clean up → RMX ligation → Beads clean up

- 500 ng input mRNA
- Ligate RTA barcode adapter (SQK-RNA002 protocol) using NEBNext Quick Ligation Module (E6056)
- RT reaction: SSIII (50 min at 50ºC), Induro (15 min 60ºC, 10 min 70ºC)
- Samples were pooled together for multiplexing
- Ligate RMX adapter using NEBNext Quick Ligation Module (E6056)

Figure 1. Direct RNA sequencing libraries were prepared following the Oxford Nanopore SQK-RNA002 protocol. 500ng mouse brain polyA (TaKaRa, cat # 636207) was used as input to make RNA/cDNA libraries with either Induro or SSIII. RNA libraries without cDNA synthesis (no_RT) were also generated. All three methods (no cDNA synthesis, or cDNA synthesis with Induro or SSIII) were used to make single-plex and multiplex libraries. Final Library yields were measured using Invitrogen™ Qubit™ dsDNA HS Assay Kit.

RESULTS

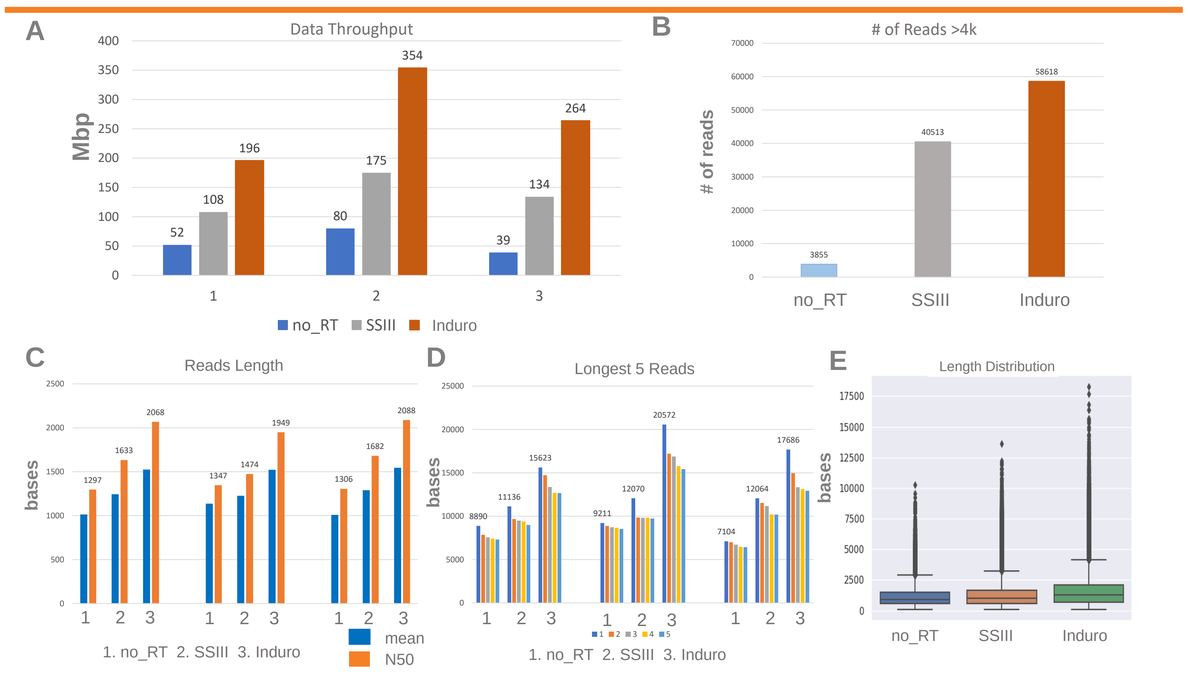


Figure 2. 3 sequencing runs were performed on pooled barcoded samples containing the direct RNA library (no_RT), the RNA/cDNA libraries made with either SSIII or Induro in the RT step. (A) Data throughput for each of the sequencing runs. The Induro libraries have significantly and consistently higher base-called data (orange bar). Qubit measurements of double stranded RNA/cDNA yield with Induro is higher than with SSIII, the ONT recommended RT (data not shown). (B) All reads from 3 multiplexed runs and two single-plex runs/sample were merged together for data analysis. The merged data was mapped to mouse mm10 genome using minimap2 . Reads (358Mb for no_RT, 3.6Gb for SSIII, 3.2Gb for Induro) were filtered for reads with length >4kb. The number of reads >4kb long is 45% higher with the Induro library (orange bar, 2.6% of total reads) than from SSIII (gray bar, 1.3% of total reads). (C) Data analysis for mean and median read lengths shows that the Induro libraries have significantly longer mean length and N50 values. The no_RT libraries have the lowest mean length and N50 values. (D) Analysis of top 5 longest reads demonstrates that Induro libraries have the longest reads up to 20kb, significantly longer than reads from SSIII and no_RT libraries. (E) Boxplot of read length distribution shows Induro libraries have longer median read lengths. Also observed is the significantly longer tail for Induro libraries, indicating more long reads.

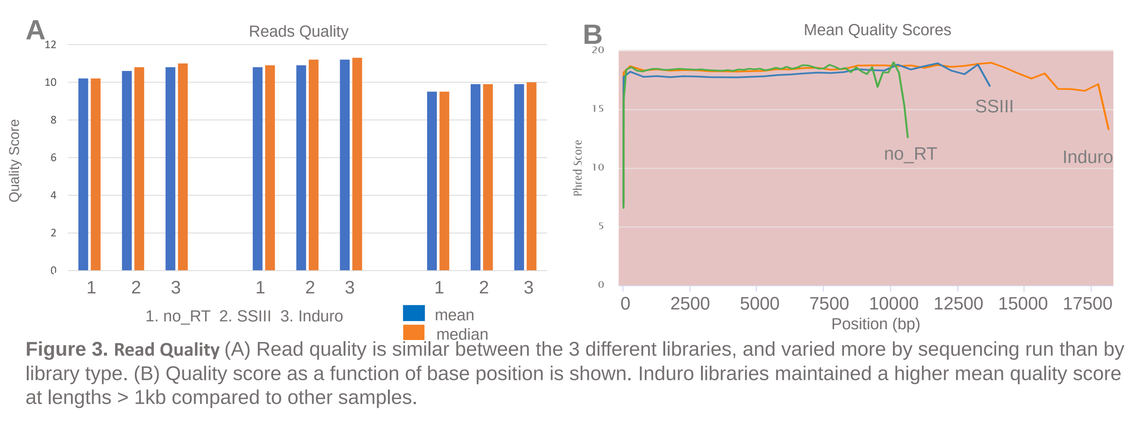


Figure 3. Read Quality (A) Read quality is similar between the 3 different libraries, and varied more by sequencing run than by library type. (B) Quality score as a function of base position is shown. Induro libraries maintained a higher mean quality score at lengths > 1kb compared to other samples.

RESULTS

no_RT
98.9% reads aligned

Read Stats		
Total reads	146,771	
- Unaligned reads	1,542	1.1%
- Aligned reads	145,229	98.9%
--- Single-align reads	144,817	98.7%
--- Gapped-align reads	35	0.02%
--- Chimeric reads	377	0.26%
----- Trans-chimeric reads	377	0.26%
----- Self-chimeric reads	0	0.00%
Base Stats *(of aligned reads)*		
Total bases	155,340,385	
- Unaligned bases	10,349,985	6.7%
- Aligned bases	144,990,400	93.3%
--- Single-aligned bases	144,775,569	93.2%
--- Other-aligned bases	14,162	0.01%

SSIII
99.3% reads aligned

Read Stats		
Total reads	314,671	
- Unaligned reads	2,093	0.7%
- Aligned reads	312,578	99.3%
--- Single-align reads	311,605	99.0%
--- Gapped-align reads	107	0.03%
--- Chimeric reads	866	0.28%
----- Trans-chimeric reads	863	0.27%
----- Self-chimeric reads	3	0.00%
Base Stats *(of aligned reads)*		
Total bases	394,326,079	
- Unaligned bases	19,964,900	5.1%
- Aligned bases	374,361,179	94.9%
--- Single-aligned bases	373,806,638	94.8%
--- Other-aligned bases	48,854	0.01%

Induro
99.5% reads aligned

Read Stats		
Total reads	506,826	
- Unaligned reads	2,710	0.5%
- Aligned reads	504,116	99.5%
--- Single-align reads	502,473	99.1%
--- Gapped-align reads	298	0.06%
--- Chimeric reads	1,345	0.27%
----- Trans-chimeric reads	1,340	0.26%
----- Self-chimeric reads	5	0.00%
Base Stats *(of aligned reads)*		
Total bases	782,448,115	
- Unaligned bases	35,776,076	4.6%
- Aligned bases	746,672,039	95.4%
--- Single-aligned bases	745,665,454	95.3%
--- Other-aligned bases	159,130	0.02%

Figure 4. Alignment Analysis All reads from 3 multiplex runs were merged for no_RT, SSIII and Induro libraries. Merged data was mapped to the mouse genome mm10 using minimap2. Mapping quality was assessed using Alignqc. Result shows very similar alignment statistics for all 3 samples. Induro and SSIII libraries have a slightly higher percentage of reads and bases aligning to the genome. Chimeric read rates were similar for all 3 libraries.

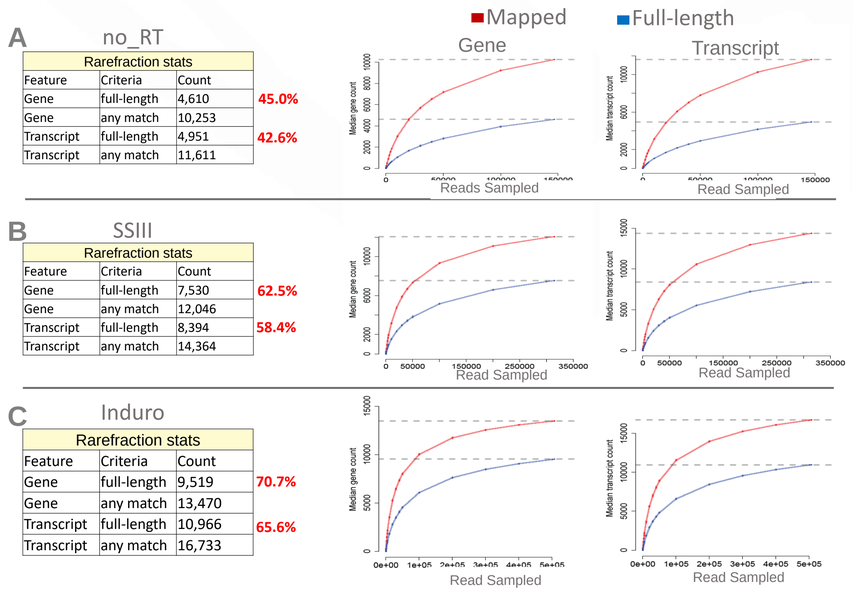


Figure 5. Annotation and rarefaction analysis. Mapped data was analyzed using Alignqc for the number of detected genes and transcripts for no_RT (A), SSIII (B), and Induro (C) libraries. The percentage of full-length genes detected is 70.7% from Induro vs 62.5% from SSIII and 45.0% from no_RT libraries. Similarly, the percentage of full-length transcripts detected was higher for Induro libraries (65.5%) compared to SSIII (58.4%) or no_RT (42.6%) libraries. This data demonstrates that higher full-length genes and transcripts were obtained with Induro.

no_RT

Alignment stats		
Best alignments sampled	901	
Base stats		
Bases analyzed	1,002,927	
- Correctly aligned bases	897,968	89.5%
- Total error bases	104,959	10.465%
--- Mismatched bases	26,282	2.621%
--- Deletion bases	49,267	4.912%
----- Complete deletion bases	28,937	2.885%
----- Homopolymer deletion bases	20,330	2.027%
--- Insertion bases	29,410	2.932%
----- Complete insertion bases	18,786	1.873%
----- Homopolymer insertion bases	10,624	1.059%

SSIII

Alignment stats		
Best alignments sampled	801	
Base stats		
Bases analyzed	1,050,043	
- Correctly aligned bases	945,093	90.0%
- Total error bases	104,950	9.995%
--- Mismatched bases	27,913	2.658%
--- Deletion bases	50,477	4.807%
----- Complete deletion bases	29,803	2.838%
----- Homopolymer deletion bases	20,674	1.969%
--- Insertion bases	26,560	2.529%
----- Complete insertion bases	16,025	1.526%
----- Homopolymer insertion bases	10,535	1.003%

Induro

Alignment stats		
Best alignments sampled	701	
Base stats		
Bases analyzed	1,113,365	
- Correctly aligned bases	1,003,544	90.1%
- Total error bases	109,821	9.864%
--- Mismatched bases	29,214	2.624%
--- Deletion bases	52,627	4.727%
----- Complete deletion bases	30,640	2.752%
----- Homopolymer deletion bases	21,987	1.975%
--- Insertion bases	27,980	2.513%
----- Complete insertion bases	16,745	1.504%
----- Homopolymer insertion bases	11,235	1.009%

Figure 6. Error statistics. A random down-sampled subset of sequencing data from no_RT, SSIII and Induro were mapped to the mouse genome mm10. Mapping data was analyzed for errors including mismatch, deletion, insertion and homopolymers using Alignqc. Similar error rates were seen with different libraries.

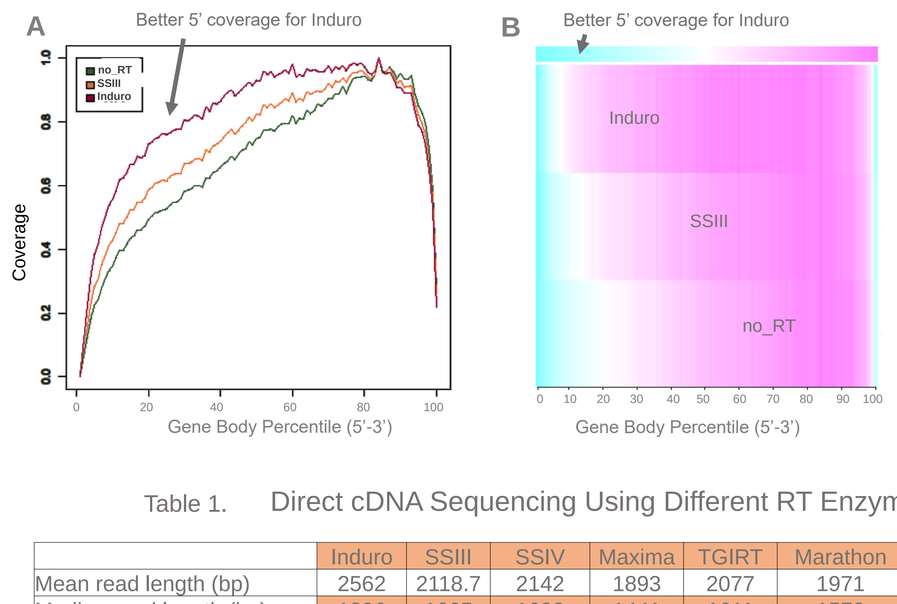


Figure 7. Gene body coverage percentile Gene Body Coverage (A) and a heat map (B) were generated using RSeQC. In the heat map, the aqua color indicates low coverage and pink indicates high coverage. In both plots, it clearly shows that Induro libraries have significantly better 5' coverage than libraries made using SSIII or made without cDNA synthesis (no_RT).

Table 1. Direct cDNA Sequencing Using Different RT Enzymes

	Induro	SSIII	SSIV	Maxima	TGIRT	Marathon	ProtoScript II
Mean read length (bp)	2562	2118.7	2142	1893	2077	1971	1995
Median read length (bp)	1896	1665	1682	1441	1611	1573	1561
Number of reads	175400	97638	148859	343925	144746	170860	149240
Read length N50 (bp)	3658	2855	2926	2681	2859	2726	2703

Table 1. Impact of RT enzymes on the length of cDNA reads following recommended conditions was studied using direct cDNA sequencing on the GridION. In this experiment, after first strand cDNA was synthesized in the RT reaction, 2nd strand cDNA was synthesized using NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module. Then the cDNA products were purified and ligated to a barcoded adapter following the protocol of Nanopore EXP-NBD104 kit. cDNA library concentrations were measured on a Qubit after the barcoding step and then equal amounts of DNA were pooled and ligated with a sequencing adapter. Finally, the resulting DNA library was cleaned up with beads and run on a GridION with a R9.4.1 flow cell. The data was demultiplexed and analyzed. Results are consistent with direct RNA sequencing data in that reads from the Induro library has the longest mean, median and N50 values.

CONCLUSIONS

- Induro, a novel reverse transcriptase shows superior performance for generating libraries for direct RNA and direct cDNA Oxford Nanopore sequencing.
- Induro is very processive, requiring a short incubation time to generate full length cDNA.
- Longer read lengths, higher percentage of full-length genes and transcripts, improved 5' coverage can be achieved with Induro compared to SSIII without increasing error rates or decreasing mapping rates.

ACKNOWLEDGEMENTS

- We thank Dr. Maggie Heider for proof reading and Dr. Guoping Ren for the discussion and assistance in developing Induro.