March 2014

University of Missouri DNA Core Facility

DNA Core Facility
216 Bond Life Sciences Center

PHONE: (573) 882-0428
FAX: (573) 884-5552
E-MAIL: dna-core@missouri.edu
http://rnet. missouri.edu/dnacore

DNA Core Evaluation of SMARTer® Ultra™ Low RNA Kit for Illumina Sequencing

Mingyi Zhou[a], Nathan J. Bivens[a], William Spollen[b;] David G. Mendoza-Cozatl[c],

[a] DNA Core Facility, University of Missouri; [b] Informatics Research Core Facility , University of Missouri; [c] Plant Sciences, University of Missouri

- Read map rates consistent with the TruSeq method
- Full-length transcipts from picogram amounts of RNA
- Consistent 5′-3′ transcript coverage

A challenge for many transciptome analysis projects is the ability to extract RNA amounts sufficient to prepare RNA-Seq libraries using standard methods (e.g. Illumina® TruSeq™ Stranded mRNA kit). Such protocols require >100 ng of total RNA as input template with the standard amount being 1-2 ug. These amounts are difficult to obtain in experiments involving only a few cells such as microdisection, cell isolation techniques, stem cells, etc.

Clontech SMARTer® template-switching technology provides a method for amplifications of small amounts of total RNA which can then be used in construction of RNA-Seq libraries. The technology produces full-length cDNA of the mRNA which becomes the input for library construction. The DNA Core Facility provide here the in-house evaluation of the SMARTer® Ultra™ Low RNA Kit for use in offering as a core service.

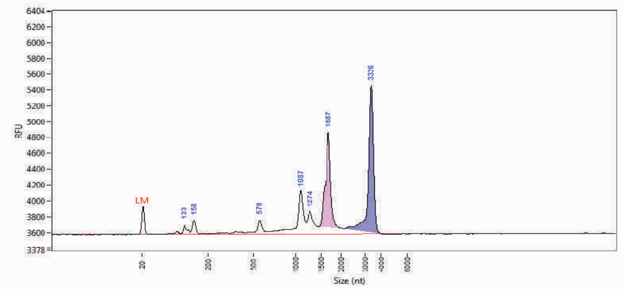


Figure 1. Electropherogram of total RNA

The Avanced Analytical Fragment Analyzer was used to evaluate the RNA integrity. Total RNA with integrity values >7.0 were used to construct RNA-Seq libraries.

Overview

Three total RNA Arabidopsis samples were provided for comparison by Dr. David Mendoza-Cozatl, Plant Sciences, University of Missouri. The DNA Core Facility confirmed RNA integrity by Fragment Analyzer to ensure minimal degradation (Figure 1). Total RNA was then prepared for library construction using the standard Illumina method. Serial dilution of the total RNA was performed to final amounts of 10 ng, 1 ng, and 100 pg for template input to the Clontech SMARTer® method. Preparation of TruSeq™ and SMARTer® libraries followed manufacturers recommendations. Sequencing was performed on a HiSeq 2500 generating between 9-45 million reads per sample with a single read, 100 base read run.   Fastq files were generated and reads demultiplexed using RTA v1.17.21.3 and CASAVA 1.8.2 (Table 1). Reads were mapped by the Informatic Research Core Facility to the reference genome Arabidopsis_thaliana.TAIR10.21 with BWA after the removal of rRNA reads. Mapped reads were summarized for gene features with total counts reported using the program featureCounts.

Sample ID	Index	Yield (Mbases)	% PF	Number Raw Reads	% Perfect Index Reads	% of >= Q30 Bases (PF)	Mean Quality Score (PF)
Sample A	GAGTGG	2,420	95.3	25,363,772	98.8	95.375	36.8
Sample A (10ng)	GCCAAT	2,052	95.1	21,566,426	96.3	94.78	36.6
Sample A (1ng)	ATCACG	2,165	94.8	22,823,232	99.2	94.71	36.5
Sample A (100pg)	CGATGT	1449	95.2	15,208,424	99.3	94.61	36.5
Sample B	AGTTCC	3,614	95.0	38,031,656	99.3	94.66	36.5
Sample B (10ng)	CTTGTA	2,196	95.0	23,116,739	99.1	94.705	36.5
Sample B (1ng)	TTAGGC	1586	94.7	16,750,208	99.3	94.295	36.4
Sample B (100pg)	ACAGTG	1666	95.0	17,514,271	99.2	94.56	36.5
Sample C	GTTTCG	4,304	95.0	45,284,374	97.8	95.335	36.8
Sample C (10ng)	TGACCA	2,227	95.0	23,423,157	99.1	94.695	36.5
Sample C (1ng)	CAGATC	872	95.1	9,164,755	98.9	94.67	36.5
Sample C (100pg)	ACTTGA	1875	95.0	19,737,862	99.1	94.66	36.5

Table 1. HiSeq run metrics

Samples were pooled in a single pool (12-plex) and loaded across 2 lanes of a HiSeq 1x100 base run. Metrics were generated with RTA v1.17.21.3 and CASAVA v1.8.2.

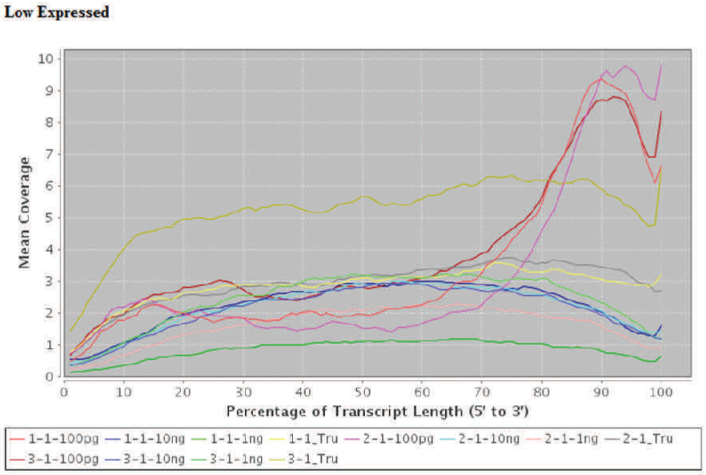
Sequence Analysis of Low RNA Input Libraries

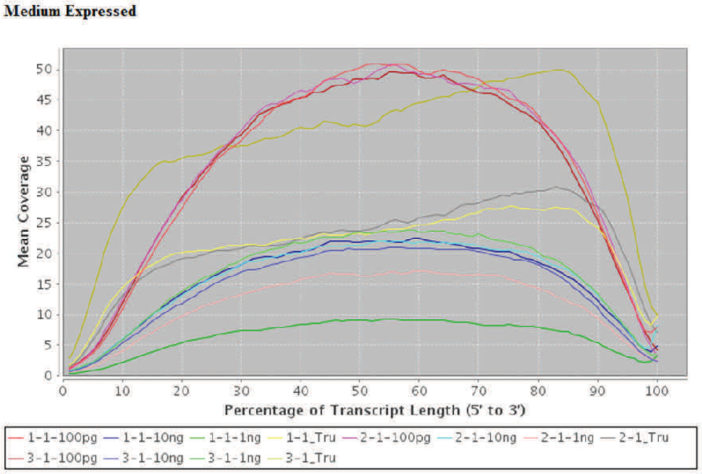
The SMARTer® Ultra™ Low RNA method involves first-strand synthesis mediated with an oligo(dT) primer following bead based enrichment for mRNA. SMARTScribe™ Reverse Transcriptase extends for the full length of the transcript and adds a few additional nucleotides which serve as a priming site for the transciptase to initiate replication of the second strand. Amplification of the full length cDNA transcripts are carried out in preparation for library construction. The use of full-length transcripts as template for library construction is expected to generate libraries with 5'-3' coverage for all transcripts with the majority of the reads mapping to the intragenic regions (within introns or exons). Therefore, libraries were analyzed with RNA-SeQC v1.1.7 (1) to measure the library metrics for comparison of Clontech SMARTer® libraries with the Illumina® TruSeq™ method.

Sample ID	Total PF Reads	Mapping Rate	Mapped Reads	Intragenic Rate	Exonic rate	Intronic Rate	Intergenic Rate
Sample A	19,685,162	0.996	19,601209	0.997	0.765	0.232	0.003
Sample A (10ng)	15,374,924	0.995	15,303,244	0.994	0.76	0.233	0.006
Sample A (1ng)	15,933,629	0.994	15,832,880	0.993	0.763	0.231	0.007
Sample A (100pg)	10,507,560	0.989	10,395,198	0.992	0.763	0.229	0.008
Sample B	19,904,528	0.999	19,889,820	0.997	0.768	0.229	0.003
Sample B (10ng)	15,675,482	0.999	15,659,407	0.994	0.762	0.232	0.006
Sample B (1ng)	11,096,304	0.998	11,078,314	0.994	0.761	0.233	0.006
Sample B (100pg)	11,695,644	0.996	11,653,380	0.993	0.767	0.226	0.007
Sample C	34,997,176	1.000	34,986,074	0.997	0.769	0.228	0.003
Sample C (10ng)	16,029,822	0.999	16,018,783	0.994	0.761	0.233	0.006
Sample C (1ng)	6,254,053	0.999	6,244,762	0.995	0.761	0.234	0.005
Sample C (100pg)	13,453,918	0.997	13,414,014	0.994	0.765	0.229	0.006

Table 2. Mapping rates of PF reads

All of the above rates are per mapped read. **Intragenic Rate** refers to the fraction of reads that map within genes (within introns or exons). **Exonic Rate** is the fraction mapping within exons. **Intronic Rate** is the fraction mapping within introns. **Intergenic Rate** is the fraction mapping in the genomic space between genes.





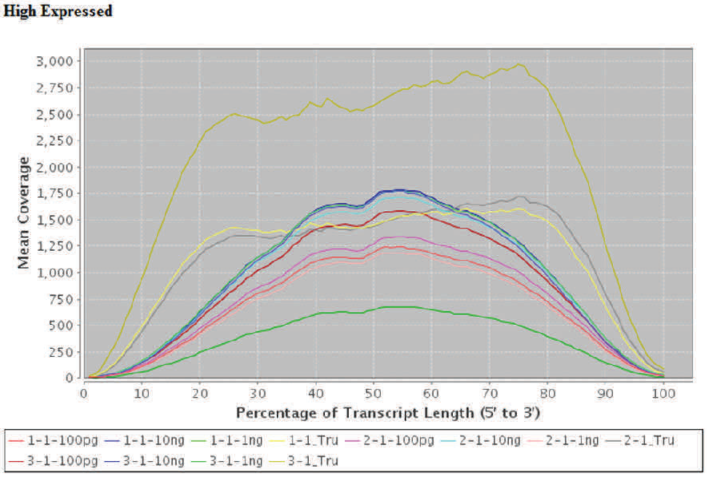


Figure 2. Mean coverage comparison of transcript length with varied amounts of Input RNA

Shown are the overlaid plots comparing average mean base coverage across the full transcript length for low, medium, and highly expressed transcripts for 2 ug (TruSeq™), 10 ng, 1 ng, and 100 pg of total RNA. The x-axis is the transcript length normalized to 100% with 0 representing the 5' end and 100 is the 3' end. Values are restricted to 1000 expressed transcripts within each expression group. 5' and 3' values are per-base coverage averaged across transcripts.

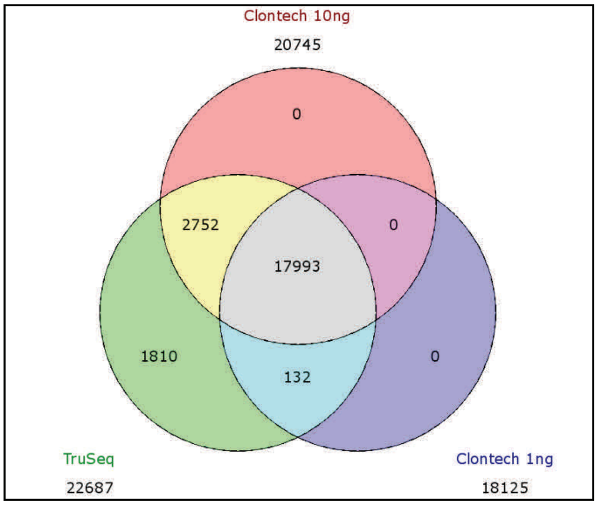


Figure 3. Venn Diagram Comparing Detected Transcripts Between Methods for Sample A.

Sample A TruSeq transcripts were sorted to remove all transcirpts with no reads detected. The transcript list was then used to compare transcripts detected with Clonetech libraries.

The overall mapping rate of reads to the reference genome was >0.993 for all samples (Table 2). The comparison of Clontech library mapping rates of PF reads was consistent with the rates identified with TruSeq™ libraries. The intragenic rate was >0.99 across all libraries and total RNA input amounts. Rates remained very low (<0.009) for reads mapping to genomic space between genes.

The TruSeq™ 5'-3' coverage when plotted for low, medium, and highly expressed transcripts illustrated the anticipated even coverage across the entire transcript length with slightly elevated coverage at the 3' end. The per base coverage average for the low, medium and highly expressed transcripts was <10, 17-50 , and 700-3000, respectively. Clontech libraries have a higher portion of reads mapping to the center of the transcript with fewer reads represented at the ends. This read distribution may be a function of the Covaris shearing method used to fragment the full length cDNA transcripts.

Studies have documented that low RNA input amounts will result in fewer genes detected; these genes typically being low expressed genes. A comparison was performed with sample A to quantify the number of transcripts identified between methods and RNA input amounts (Figure 3). A total of 22,687 unique transcripts were identified for sample A using the TruSeq method. As expected, the total RNA input amount of 1ng identified significantly fewer transcripts (4,562 lacking a single read). The 10ng input amount was near similar to the TruSeq™ method with only 132 transcripts not represented.

Reproducibility of the Clontech SMARTer® method, especially in comparison to results previously obtained by researchers with the TruSeq™ method, was of importance to the DNA Core in determining. Read counts were normalized across libraries to allow for comparison. For direct comparison of the same transcripts, all transcripts with no reads were removed from the TruSeq™ dataset. Corresponding transcripts from the Clontech SMARTer® library, with normalized read counts, were then matched to the TruSeq™ dataset to compare the reproducibility of the low input libraries. Scatter plots were generated to compare read counts for each sample between methods and RNA input amounts (Figure 4). The results demonstrate a high correlation of transcript levels between the TruSeq™ method and the Clontech SMARTer® method at 10 and 1 ng. The 100 pg input amount, however, shows much more variability that may be improved upon with protocol optimization such as increased PCR cycles when amplifying the full length cDNA.

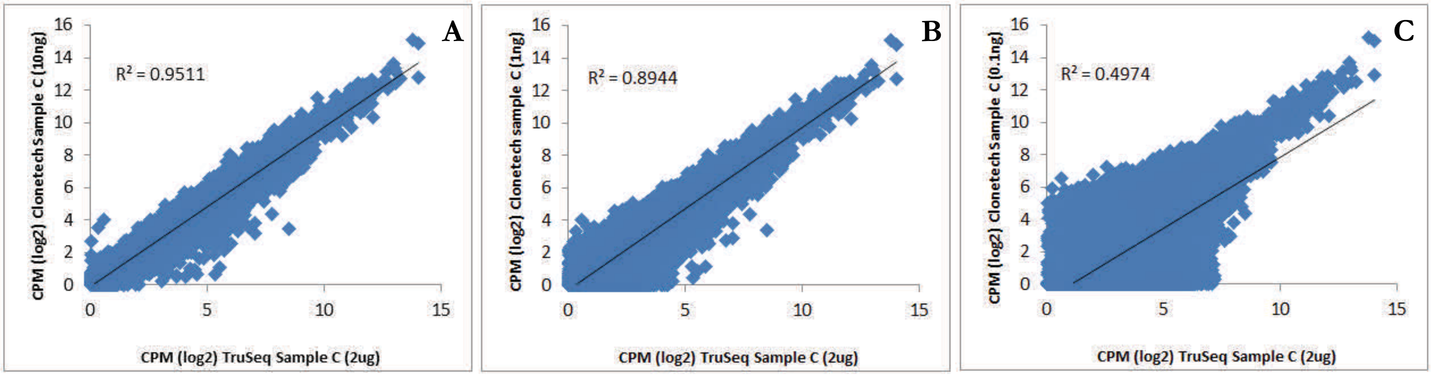


Figure 4. Reproducibility of read counts compared between Illumina® TruSeq™ and SMARTer® Ultra™ Low RNA Kit

Scatter plots compare the reproducibility of transcript read counts of the SMARTer® method with low RNA input to that of the standard TruSeq™ method. Sample C read counts were normalized to counts per million across all libraries. The normalized read counts were consistently reproduced with the 10 ng and 1 ng RNA inputs (Panels A and B). When the RNA input was lowered to 100 pg the reproducibility decreased significantly (Panel C).

Summary

The SMARTer® Ultra™ Low RNA kit provides a reproducible method for the amplification of mRNA from sub-nanogram RNA input amounts that is suitable for the production of Illumina sequencing libraries. Libraries constructed from the full-length cDNA transcripts are shown to produce sequences that map at rates consistent with the TruSeq™ method which use total RNA starting amounts of 1-2 ug. The analysis of the intragenic mapping rate (exon + intron coverage) has shown to be equivalent to the TruSeq™ method. The SMARTer® method also provides reproducibility in total transcripts detected down to 1 ng of total RNA as demonstrated with a high correlation in normalized read counts. In addition, the 5'-3' coverage demonstrates the effectiveness of the SMARTScribe™ Reverse Transcriptase to produce full-length transcipts from picogram amounts of RNA. These results are consistent with previous studies comparing commercially available kits (2).

In summary, the SMARTer® Ultra™ Low RNA kit provides a robust method for preparation of Illumina sequencing libraries from nanogram amounts of total RNA that yield reproducible RNA-seq data.

References

1. DeLuca, David S., *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28(11), 1530-1532 (2012).

2. Shanker, S., *et al.* Evaluation of Commercially available RNA Amplification Kits at Subnanogram Input Amounts of total RNA for RNA-seq. 2013. *ABRF Study*