Evaluation of RNA sequencing methods for use with Highly degraded Formalin-Fixed, Paraffin-embedded (FFPE) tissue samples.

Julie A. Dragon[1,2], Ramiro Barrantes-Reynolds[1,2], Jessica Hoffman[1] and Scott Tighe[1].

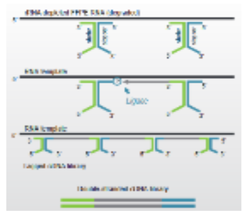
[1] Vermont Integrative Genomics Resource, [2] UVM LCOM Department of Microbiology and Molecular Genetics

Introduction

Genomics research requires a significant sample size to provide robust biological signal, leaving researchers clamoring for access to large sample sets. As this research continues to expand into the clinical arena, the demand to sequence RNA from banked tissue samples, such as formalin-fixed paraffin-embedded blocks, is unavoidable. Recovering DNA and RNA from such samples can be challenging depending on age of the sample block and fixture protocols. To fulfill the need for increased recovery of usable reads, several manufacturers have developed solutions to address these challenges, including FFPE-specific extraction kits, as well as library synthesis and quality control reagents. **In this study, we analyzed the quality and outcome of RNA-Seq data generated from three library synthesis kits of FFPE-derived human thyroid tumors with storage times from 3-6 years.**

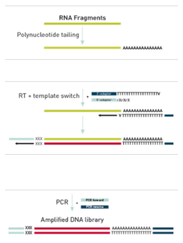
X Total RNA Library Prep Kit V1 4-10-16

- Ribosomal depletion is critical
- Rapid, high efficiency stranded library prep with strand-displacement stop/ligation technology
- RNA degradation down to DV200 of 44%
- Not specified for FFPE



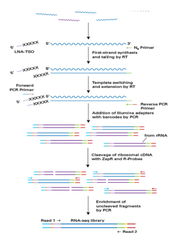
Diagenode CATS Total RNA-seq Kit V2I [with rRNA depletion]

- Total RNA "Capture and Amplification by Tailing and Switching" (CATS)
- Ligation free, high efficiency method, inputs down to 100 pg
- Single tube, ultra fast library prep, maintains strandedness
- Specified for FFPE but unknown performance specs



Takara SMARTer Strantded Total RNA-Seq Kit V2 - Pico Input Mammalian

- Total RNA SMARTer technology (Switching Mechanism At 5' end of RNA Template) Using Locked Nucleic Acid technology
- Ultra low degraded RNA input, with RIN quality down to 3
- Rapid rRNA depletion with proprietary Zap probes, strandedness



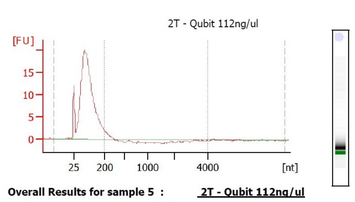
Methods

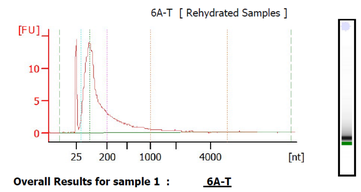
RNA was extracted in 2016 from all samples using the QIAGEN All Prep FFPE Kit. SE100 sequencing was performed on the Illumina HiSeq1500. The following table and bioanalyzer traces describe features of the samples analyzed to date and the three kits used to prepare the libraries:

Table 1. Thyroid tumor samples derived from FFPE blocks. Samples colored by age and ordered by tumor type; T = Indolent tumor and N = its associated normal tissue, and AT = Aggressive tumor and AN = its associated normal tissue. Sample IDs colored by year fixed. Red to white coloring indicates better to worse values for each quality metric.

Sample ID*	Year fixed	Patient age	miRNA array image	%BA >200
6AT	2009	51	5	31
11AN	2010	52	4	33
2T	2012	32	3.5	<10
5T	2011	42	3	40
4N	2012	44	4	47
6N	2009	58	5	32

Fig. 1. Bioanalyzer traces for two samples from the sample set.





The data processing pipeline included adaptor and poor quality base trimming (trimommatic), removal of rRNA reads (sortMeRNA), genomic alignment to hg19 (HiSat2), and read distribution and alignment quality assessment (RNASeqQC) (Adiconis et al, 2013).

Adiconis, X., D. Borges-Rivera, R. Satija, D. S. DeLuca, M. A. Busby, A. M. Berlin, A. Sivachenko, et al. 2013. "Comprehensive comparative analysis of RNA sequencing methods for degraded or low input samples." Nature methods 10 (7): 10.1038/nmeth.2483. doi:10.1038/nmeth.2483.

Results

Results indicated that of the kits tested, it is possible to capture usable RNA-Seq data from highly degraded FFPE tissues. Features such as age of block, RNA fragment length, crosslink time exposure, and read duplication are important considerations. However, it is clear that quality assessment requires multiple metrics, including number of mappable reads, % rRNA, % duplication, and % exonic reads.

Sample	Fixation Year	Total # of Reads	rRNA Reads	% of rRNA	Total # of Filtered Reads	# Reads Mapped Unique	Mapped Unique Rate of Total	Duplication Rate	Exonic Rate	Intronic Rate	Expression Profiling Efficiency	Transcripts Detected
6AT_Diagenode	2009	19,908,984	**6,219,525**	31%	13,689,459	1,407,694	0.103	**0.639**	0.203	0.723	0.058	27,305
6AT_X		44,471,163	16,927,525	38%	26,988,422	377,866	0.014	0.975	0.352	0.545	0.199	27,041
6AT_Takara		**80,483,884**	21,325,354	27%	**59,158,530**	**12,115,956**	**0.205**	0.747	**0.353**	**0.521**	**0.285**	**39,412**
11AN_Diagenode	2010	31,992,774	4,740,647	15%	27,252,127	**8,263,951**	0.303	**0.307**	0.244	0.687	0.107	35,467
11AN_X		**56,044,722**	**19,770,719**	35%	**34,109,506**	30,619	0.001	0.998	**0.256**	**0.624**	0.117	2,870
11AN_Takara		17,138,050	**2,914,672**	17%	14,223,378	6,113,065	**0.430**	0.49	0.248	0.672	**0.209**	**36,580**
2T_Diagenode	2012	4,814,741	3,139,128	65%	1,675,613	372,116	**0.222**	**0.172**	0.146	0.745	0.039	7,821
2T_X		**38,058,432**	**1,781,668**	5%	**35,364,964**	73,735	0.002	0.997	0.175	0.694	0.115	7,491
2T_Takara		30,125,563	6,314,581	21%	23,810,982	**4,540,022**	0.191	0.653	**0.226**	**0.692**	**0.124**	**35,082**
5T_Diagenode	2011	3,406,909	2,897,324	85%	509,585	50,210	0.099	**0.518**	0.155	**0.376**	0.032	1,613
5T_X		**35,302,099**	**1,094,941**	3%	**34,339,262**	46,559	0.001	0.996	0.158	0.483	0.059	4,712
5T_Takara		9,201,164	2,422,335	26%	6,778,829	**858,172**	**0.127**	0.807	**0.309**	0.583	**0.202**	**29,031**
4N_Diagenode	2012	25,207,587	6,656,598	26%	18,550,989	2,478,779	0.134	0.597	0.219	0.703	0.073	31,823
4N_X		**31,880,985**	9,727,165	31%	**19,865,785**	40,992	0.002	0.996	0.226	**0.608**	0.132	4,910
4N_Takara		21,813,304	**3,168,621**	15%	18,644,683	**8,035,052**	**0.431**	**0.490**	**0.261**	0.642	**0.221**	**37,505**
6N_Diagenode	2009	22,399,649	4,966,355	22%	17,433,294	1,504,217	0.086	0.604	0.144	0.799	0.031	25,761
6N_X		**36,970,368**	4,755,142	13%	**32,127,658**	76,977	0.002	0.997	0.262	0.660	0.199	10,177
6N_Takara		23,650,220	**3,763,245**	16%	19,886,975	**7,959,531**	**0.400**	**0.493**	**0.268**	**0.655**	**0.212**	**36,352**

Table 2. RNA sequencing quality metrics. Samples colored by year of fixation. Metrics are colored gold to light yellow from highest to lowest, and preferred numbers for each metric are in bold.

By most metrics and regardless of kit used, we identified a clear "best" and "worst" sample, 6AT and 5T, respectively, with the remaining samples within this range. **Age was not a determining factor in this case**, as the best and worst samples were the oldest in the set (2009).

While most kits performed to manufacturers specifications, the Takara kit seems to be the least influenced by severe RNA degradation with higher **uniquely mapped rate, exonic read coverage, and transcripts represented**, even with the most challenging sample (Table 2).

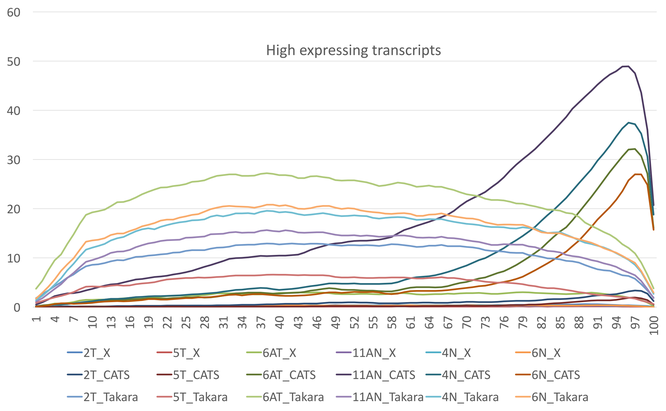


Fig 2. Mean normalized coverage by position of all samples for high expressing transcripts. *Mean Coverage* is the y-axis, and *Percent of Transcript Length (5' to 3')* is the x-axis.

High and low expressing transcripts were more evenly covered at about 30x in the Takara prepared samples, while the CATS prepared samples showed higher (~50x) coverage predominantly at the 3' end of the transcripts (Fig. 2, all samples, Fig. 3, "best and worst" only). The X prepared samples had an even distribution of very low coverage across transcripts

Both Diagenode and Takara kits had lower duplication rates than kit X, with Diagenode lower than Takara. The most significant issue for the X sample preparations seems to be the duplication rate, which is likely due to the high number of PCR cycles needed to amplify the low input libraries. However, this was a kit adapted for early use on FFPE, not specifically design for such samples.

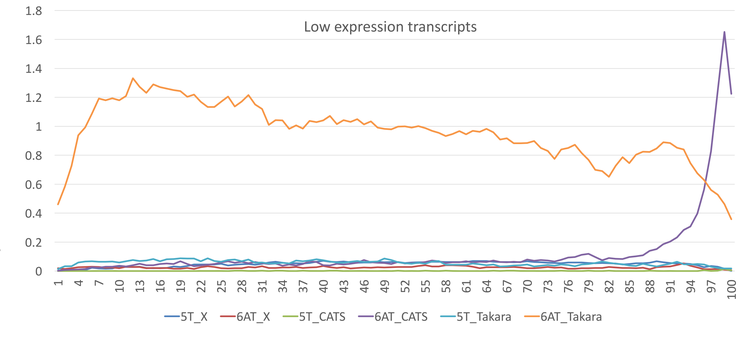


Fig 3. Mean normalized coverage by position of the "best and worst" samples for low expressing transcripts. *Mean Coverage* is the y-axis, and *Percent of Transcript Length (5' to 3')* is the x-axis.

Conclusions and Future Directions

Our analysis illustrates some of the features of the tested synthesis kits that enable one or another to perform better for variable quality FFPE samples. Low duplication and efficient ribo-depletion are key features for highly degraded sample prep kits. Future work include the addition of more samples, as well as metadata on the preparation of each sample (ie time from resection to fixation, time in formalin, and side of block sampled).

Acknowledgments

We would like to thank the the UVM Larner College of Medicine, the Lake Champlain Cancer Research Organization, and the UVM Medical Center for their continued support.