Date: December 2nd, 2019

Department of Genomic Medicine
Cancer Immune Monitoring and Analysis Center
Division of Cancer Medicine
The University of Texas MD Anderson Cancer Center

Director:
Andy Futreal, BS, Ph.D., Professor and Chair
Curtis Gumbs, Scientific Manager
Jianhua (John) Zhang, Ph.D., Director, Computational Genomics, Genomic Medicine

RNA Sequencing Analytical Validation

Version 3.0

This report describes the analytical validation parameters for RNA sequencing assay performed at MD Anderson Cancer Center CIMAC.

RNA Sequencing	
Analytical Performance	Quality control (QC) was performed on sample nucleic acid and library preparations prior to sequencing. The total number of reads, genes detected, exonic, intronic and intergenic rates were calculated. The mean number of reads for FFPE and FF tissues was 169.53 and 191.80 respectively. Per sample analysis demonstrated high reproducibility and consistency in the total number of reads, detected genes and exonic, intronic and intergenic rates in both FFPE and FF specimens across triplicates and runs.
Analytical Reproducibility	Analytical reproducibility was evaluated within a run, across run and between time points (inter-run) for each sample by clustering analyses. Each patient sample clustered together showing the reproducibility of the RNA seq generated data.
Analytical Sensitivity	The analytical sensitivity of the sequencing data was determined by the total number of genes detected per sample triplicate and runs. A high concordance in a number of genes was observed. Additionally, Differentially Expressed Genes (DEGs) for RNA-seq between tumor and normal (blood) across triplicates and sample type were calculated (FFPE and FF).
Any other performance characteristics required for assay performance	All of the required equipment have annual service contracts with regular Preventive Maintenance performed to maintain optimal calibration and performance. All other small equipments such as multi-channel pipettes and laboratory material have calibration performed by certified vendors.

The analytical validation assay was performed in two steps. In the first step, the inter and intra-assay reproducibility was performed using 3 lung tumor cases with Fresh Frozen (FF), Formalin-Fixed Paraffin-Embedded (FFPE) tissues and Peripheral Blood Mononucleated Cells (PBMCs). One sample showed poor tumor content and was eliminated from further processing. In the second step a pilot experiment concordance between FF and frozen tissue was analyzed with addition 4 tumor cases to extend the number of samples.

Step 1

1. **Samples:** Fresh Frozen (FF), Formalin-Fixed Paraffin-Embedded (FFPE) tissues and Peripheral Blood Mononucleated Cells (PBMCs) representing three lung cancer cases (MDA-5760: Adenocarcinoma, MDA-5812 and MDA-5971: Squamous Cell Carcinoma) were chosen for the analytical validation.

   Sample Quality Control (QC): Histological and cytological examination and assessment of FF and FFPE tissue specimens were done by a reference pathologist. Tissue quality was assessed before the extraction of RNA. All H&E stained histological samples used for QC were scanned and digital images are available for review.

2. **RNA extraction:** RNA from the samples representing the three patients was isolated by the following methods.
   (i) RNA was extracted from PBMCs and FF tissues using the Qiagen miRNeasy Mini Kit (cat.217004).
   (ii) RNA was extracted from available FFPE tissues using FFPE tissue Qiagen RNAeasy FFPE kit (cat. 73504).

   RNA Quality Control (QC): Total RNA was first assessed using a spectrophotometric method to determine concentration. The 260/280 ratio was within 1.8 - 2.0 for all samples (**Table1**). The Ab_{260}/Ab_{230} or Ab_{260}/Ab_{280} ratio >= 1.8 was kept as the cutoff for quality check. Total RNA was further analyzed on the Agilent TapeStation 4200. DV_{200} values for all FF tumors were above 80% and FFPE and PBMC were between 55-70%.

Table 1. Study samples Information and quality control.

**Tumor content (%): Percentage of tumor in tissue; **Malignant cell (%): Percentage of viable tumor cells; ***MDA-5812-T, a sample without tumor; N/A: not applicable*

MDA Sample ID	Tissue Type	RNA conc (ng/ul)	% Tumor	% Malignant Cell	RNA Integrity Number (RIN)	RNA Quality (Ab_{260}/Ab_{280} ratio)	DV200 (%)	Diagnosis
MDA-5760-T	**Frozen**	397.7	70	70	**9.2**	2.06	86.2	Adenocarcinoma
MDA-5760-fpT	**FFPE**	382.0	90	90	**2.4**	2.05	54.8	Adenocarcinoma
MDA-5760-C	**PBMC**	59.4	N/A	N/A	**7.5**	1.83	69.2	Adenocarcinoma
MDA-5812-T	**Frozen**	166.7	0	0	**7.4**	2.02	80.7	Squamous
MDA-5812-fpT	**FFPE**	484.7	80	60	**2.3**	2	63.8	Squamous
MDA-5812-C	**PBMC**	83.0	N/A	N/A	**6.9**	1.9	70.3	Squamous
MDA-5971-T	**Frozen**	668.6	60	20	**8.3**	2.09	85.9	Squamous
MDA-5971-fpT	**FFPE**	181.1	40	40	**2.3**	2.03	60.8	Squamous

Further, using the grading criteria shown in **Table 2**, all the FFPE samples were ranked good quality for RNA sequencing-based on their DV_{200} values.

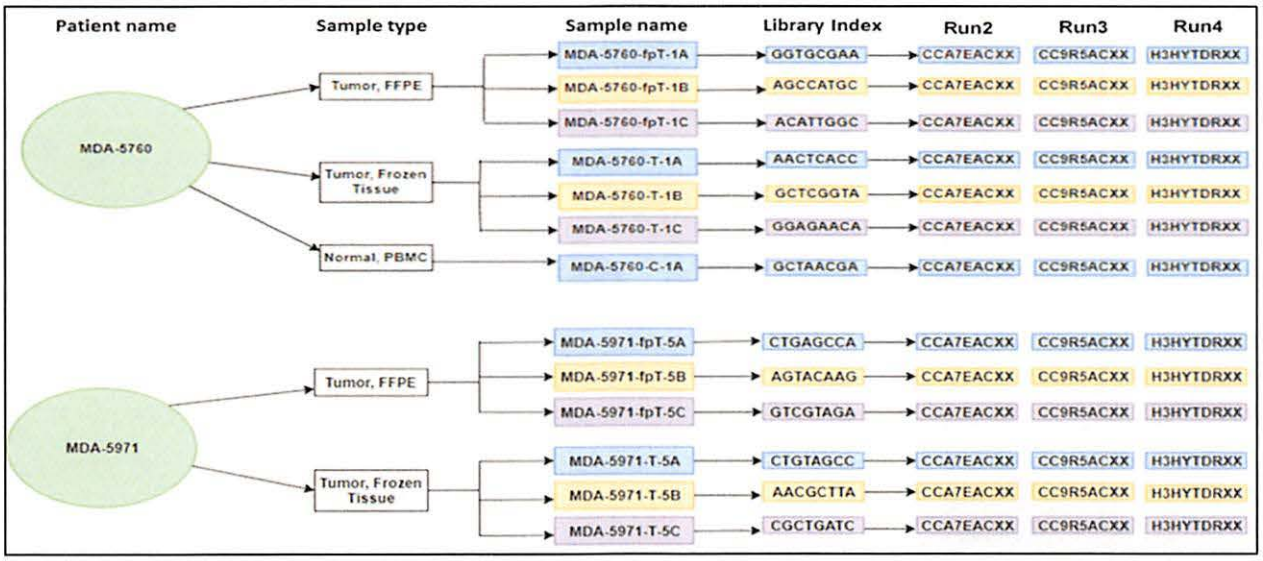
Table 2. Grading criteria of RNA samples based on the percentage of RNA greater than 200bp in length

Grade	Percentage of RNA >200 nt
Good FFPE RNA	>50%
Poor FFPE RNA	20% to 50%
Inapplicable FFPE RNA	<20%

3. Study Design.

Due to the low tumor content of MDA-5812, it was not included in further processing. Details of the RNA-Seq workflow at MDACC are described in Supplemental Figure 1. RNA sequencing was performed on RNA extracted from FF, FFPE and PBMC samples corresponding to two lung cancer cases (one adenocarcinoma, two squamous cell carcinoma) selected at MDACC by using a 2x3x3 experimental design. RNA isolated from the six samples (**Table 1** and **Figure 1 Column 2**) were used for library generation in triplicate (6X3 libraries). Thus, the same RNA was used to generate the triplicates. The 18 libraries generated from 6 samples were sequenced 3 times independently. Details of study samples and experimental study design are shown in **Table 1** and **Figure 1.**

Figure 1. Experimental study design.



Library Preparation and sequencing: After passing RNA quality check, 20 ul of RNA at a concentration of 20 ng/ul were aliquoted for library preparation and sequencing. Although RIN values were low for the FFPE samples, the >50% DV_{200} values justified their inclusion in the studies.

Individual RNA libraries were prepared using the Agilent SureSelect^{XT} RNA direct protocol and sequenced using the Illumina paired-end sequencing platform (SureSelect Human All Exon Kit V6). Illumina's HiSeq 2500 SN 1222 and NovaSeq SN A00482 sequencing platforms were used to do this sequencing (Run 1, HiSeq; Run 2 and 3, Nova Seq). Sequencing length of PE76, depth of 100M paired ends. The quality of the raw data conformed to Hiseq and NovaSeq's standards.

Library Quality Control (QC):

Prehybridization library QC. After the total RNA samples were converted to cDNA prehybridization libraries, samples were analyzed using the Agilent TapeStation 4200. The criteria set was that the electropherogram for each sample should demonstrate a peak positioned between 150 to 450 bp. Hybridization requires 200 ng of each prehybridization library. Any sample that fails to generate more than 200 ng of prehybridization library or fails to demonstrate a peak in the noted range is considered a failed sample.

Final library QC. After hybridization and final PCR amplification, samples were analyzed using the Agilent TapeStation 4200. The criteria set was that the electropherogram for each sample should demonstrate a peak positioned between 200 to 500 bp. Any sample that fails to generate more than 5 nM of the final library, fails to demonstrate a peak in the noted range or does not generate a CT value above 2 nM via qPCR is considered a failed sample.

Data Analyses and Bioinformatics.
BCL (raw output of HiSeq) files were processed using Illumina's CASAVA (Consensus Assessment of Sequence And Variation) tool for de-multiplexing/conversion to FASTQ format, which is the standard input for most aligners and downstream analytic tools.
The FASTQ files were aligned to the reference genome (human Hg19) using STAR (Dobin et. al. 2012) following the two-step alignment procedure.
The generated BAM files are subject to the quantification of gene expression using HTSeq (Anders et. al. 2014)

Analytical Performance.
Duplicate Reads and Number of Detected Genes. The basic sequencing metrics were evaluated including the total number of reads, duplicate reads of mapped reads, number of detected genes and percentage of reads in different genomic regions (exonic, intronic and intergenic). **Table 2** shows the total number of reads observed per triplicate and per sample type. As expected a higher number of reads were observed in FF samples (total reads FF vs FFPE; 191.80 vs 169.53).

Table 2. RNA-Seq total reads (in millions) in FFPE and FF tumors. (Sample ID consists of 4 parts: NGS run number-patient-sample type-Replicate ID).

Sample ID	FFPE (fpT)					FF (T)				
	Run 2	Run 3	Run 4	Mean	Sample Mean	Run 2	Run 3	Run 4	Mean	Sample Mean
MDA-5760-1A	137.99	163.29	154.41	**151.90**		233.83	221.85	204.85	**220.18**	
MDA-5760-1B	217.68	193.42	180.76	**197.29**	**176.45**	163.62	183.63	169.11	**172.12**	**189.28**
MDA-5760-1C	199.58	175.94	164.94	**180.15**		168.63	185.99	172.04	**175.55**	
MDA-5971-5A	158.53	165.71	155.84	**160.03**		210.98	213.15	196.58	**206.90**	
MDA-5971-5B	143.61	172.3	162.15	**159.35**	**162.61**	188.48	227.79	211.26	**209.18**	**194.31**
MDA-5971-5C	149.29	183.56	172.54	**168.46**		172.47	170.56	157.56	**166.86**	
				TOTAL FFPE MEAN	**169.53**				**TOTAL FF MEAN**	**191.80**

The duplicate rate of mapped reads was plotted by library preparation per sample (**Figure 2A**) and per sample by run (**Figure 2B**). Duplication rate is always high in RNA-seq due to the enrichment of reads covering the transcripts. Analyses of the sample showed high reproducibility and consistency in the duplicate rates in both FFPE and FF specimens across triplicates.

Figure 2A. Duplicate rates of mapped reads (each bar represents a patient and sample types are color-coded).

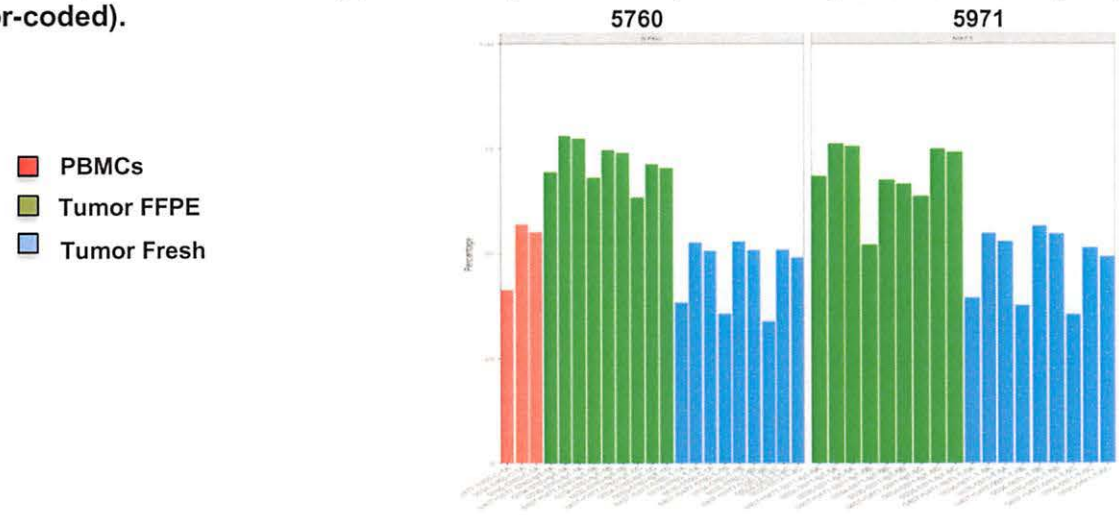
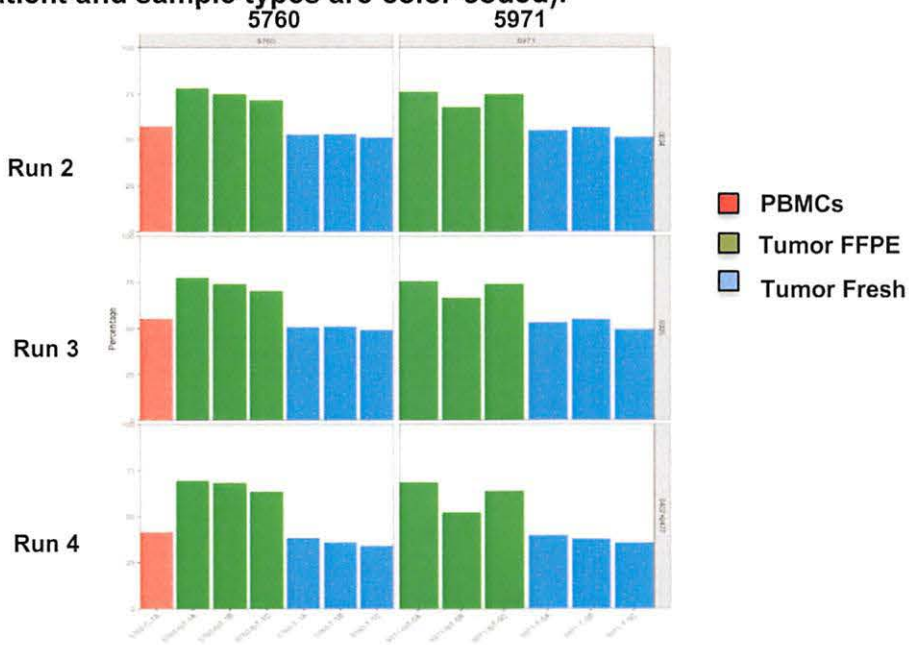


Figure 2B. Duplicate rates of mapped reads per replicate and run (each bar represents a patient and sample types are color-coded).



A total number of genes detected per sample replicate by individual sequencing run shows an FFPE mean of 25068.50 and FF Mean 24816.96 genes (**Table 3** and **Supplemental Figure 2A and 2B**). A high number of genes were detected across samples and runs showing good analytical performance and reproducibility.

Table 3. The number of detected genes in FFPE and FF tumors.

Sample ID	FFPE (fpT)					FF (T)				
	Run 2	Run 3	Run 4	Mean	Sample Mean	Run 2	Run 3	Run 4	Mean	Sample Mean
MDA-5760-1A	24003	24261	24173	**24145.67**	**24801.00**	25315	25163	25016	**25164.67**	**24742.78**
MDA-5760-1B	25384	25149	24957	**25163.33**		24309	24582	24481	**24457.33**	
MDA-5760-1C	25365	25017	24900	**25094.00**		24489	24688	24642	**24606.33**	
MDA-5971-5A	24948	25006	24950	**24968.00**	**25336.00**	25096	25096	24954	**25048.67**	**24891.11**
MDA-5971-5B	25487	25865	25694	**25682.00**		24812	25231	25110	**25051.00**	
MDA-5971-5C	25141	25508	25425	**25358.00**		24687	24626	24408	**24573.67**	
TOTAL FFPE MEAN					**25068.50**	TOTAL FF MEAN				**24816.96**

A similar percentage of reads in the different mapped regions was observed across sample triplicates and runs (**Figure 4A-4C and Supplemental Figure 3**)

Table 4A. The fraction of reads in exonic regions.

Sample ID	FFPE (fpT)					FF (T)				
	Run 2	Run 3	Run 4	Mean	Sample Mean	Run 2	Run 3	Run 4	Mean	Sample Mean
MDA-5760-1A	0.822	0.819	0.818	**0.819**	**0.817**	0.882	0.881	0.880	**0.881**	**0.880**
MDA-5760-1B	0.820	0.817	0.816	**0.818**		0.882	0.882	0.880	**0.881**	
MDA-5760-1C	0.818	0.814	0.814	**0.815**		0.879	0.879	0.878	**0.879**	
MDA-5971-5A	0.821	0.818	0.818	**0.819**	**0.818**	0.861	0.861	0.859	**0.860**	**0.860**
MDA-5971-5B	0.820	0.818	0.817	**0.818**		0.862	0.861	0.860	**0.861**	
MDA-5971-5C	0.819	0.817	0.816	**0.817**		0.858	0.857	0.856	**0.857**	
TOTAL FFPE MEAN					**0.8175**	TOTAL FF MEAN				**0.870**

Table 4B. The fraction of reads in intronic regions.

Sample ID	FFPE (fpT)					FF (T)				
	Run 2	Run 3	Run 4	Mean	Sample Mean	Run 2	Run 3	Run 4	Mean	Sample Mean
MDA-5760-1A	0.123	0.122	0.123	**0.123**	**0.123**	0.087	0.086	0.087	**0.087**	**0.087**
MDA-5760-1B	0.123	0.122	0.122	**0.122**		0.086	0.085	0.086	**0.086**	
MDA-5760-1C	0.124	0.122	0.123	**0.123**		0.088	0.087	0.088	**0.088**	
MDA-5971-5A	0.131	0.130	0.131	**0.131**	**0.133**	0.103	0.102	0.103	**0.102**	**0.102**
MDA-5971-5B	0.135	0.134	0.135	**0.135**		0.101	0.100	0.101	**0.101**	
MDA-5971-5C	0.135	0.134	0.134	**0.134**		0.103	0.102	0.103	**0.103**	
TOTAL FFPE MEAN					**0.128**	TOTAL FF MEAN				**0.095**

Table 4C. The fraction of reads in intergenic regions.

Sample ID	FFPE (fpT)					FF (T)				
	Run 2	Run 3	Run 4	Mean	Sample Mean	Run 2	Run 3	Run 4	Mean	Sample Mean
MDA-5760-1A	0.054	0.058	0.058	**0.057**	**0.059**	0.030	0.031	0.031	**0.031**	**0.031**
MDA-5760-1B	0.056	0.060	0.061	**0.059**		0.030	0.031	0.031	**0.031**	
MDA-5760-1C	0.058	0.062	0.062	**0.061**		0.031	0.032	0.032	**0.031**	
MDA-5971-5A	0.047	0.050	0.051	**0.049**	**0.048**	0.035	0.037	0.037	**0.036**	**0.038**
MDA-5971-5B	0.044	0.047	0.047	**0.046**		0.036	0.038	0.038	**0.037**	
MDA-5971-5C	0.045	0.048	0.049	**0.047**		0.038	0.040	0.040	**0.039**	
TOTAL FFPE MEAN					**0.054**	TOTAL FF MEAN				**0.0345**

4. Analytical Reproducibility:

Within-run and across-run reproducibility. Normalized Transcripts per Kilobase Million (TPM) values were log-transformed and then Z-score scale data obtained to determine pairwise alignment. The pairwise alignment of all genes across all sample triplicates of FF, FFPE tumors and normal samples show good intra-run and inter-run consistency of the data (**Figure 5A**). The samples are labeled as sample ID.tumor/normal.run #. Sample IDs represent Patient number.tissue-type.replicate ID as shown in **Figure 1.**

Inter-run reproducibility. Inter-run reproducibility was assessed by using sequencing data of the same samples obtained three months prior to the current study (Run1). **Table 5** shows analytical performance metrics of this prior run including total reads, number of detected genes and percentage of reads in intronic, exonic and intergenic regions from that sequencing. Samples from run 1 were not run in replicates and hence do not have replicate IDs.

Figure 5A. Clustering analyses of FF, FFPE tumors and PBMC. Sample Label are as follows
5760.fpT.1A.T.2 - Patient number.tissue type(FFPE/FF/PBMC).replicate ID. tumor/normal.run #

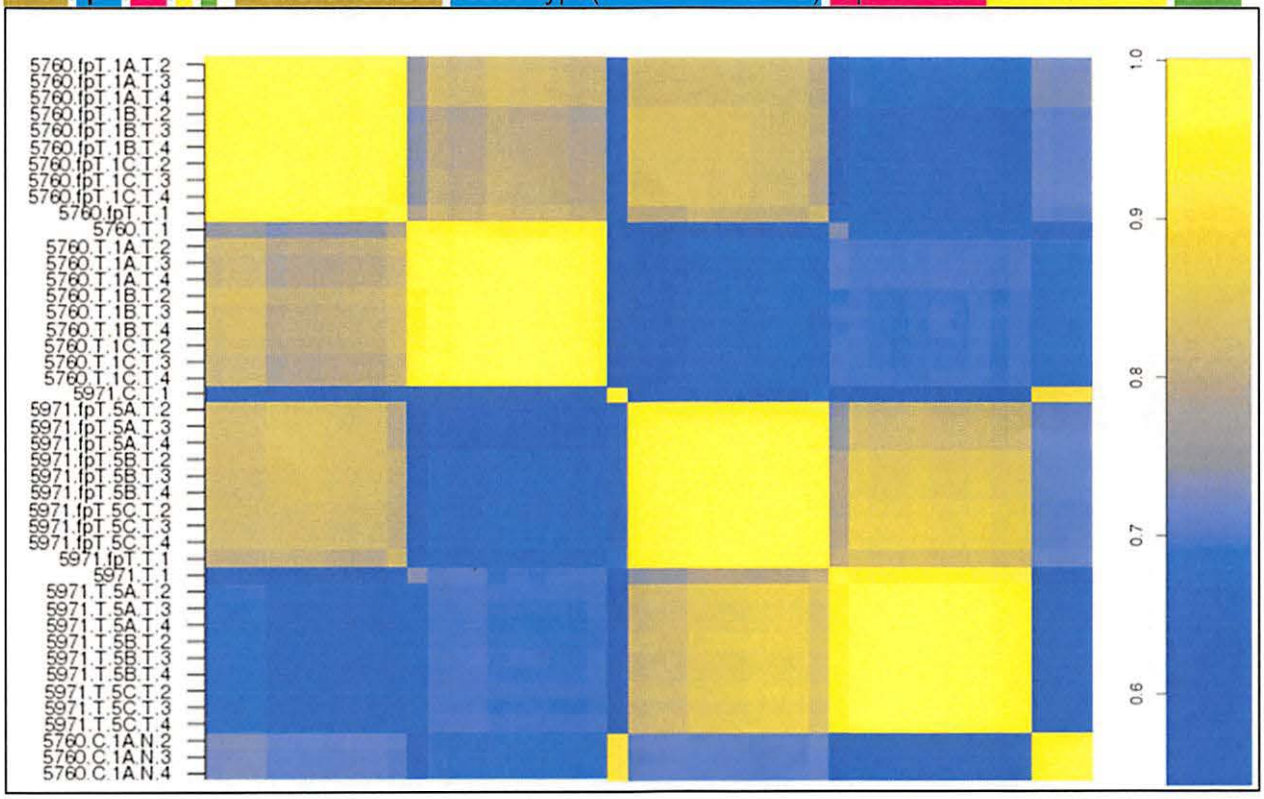
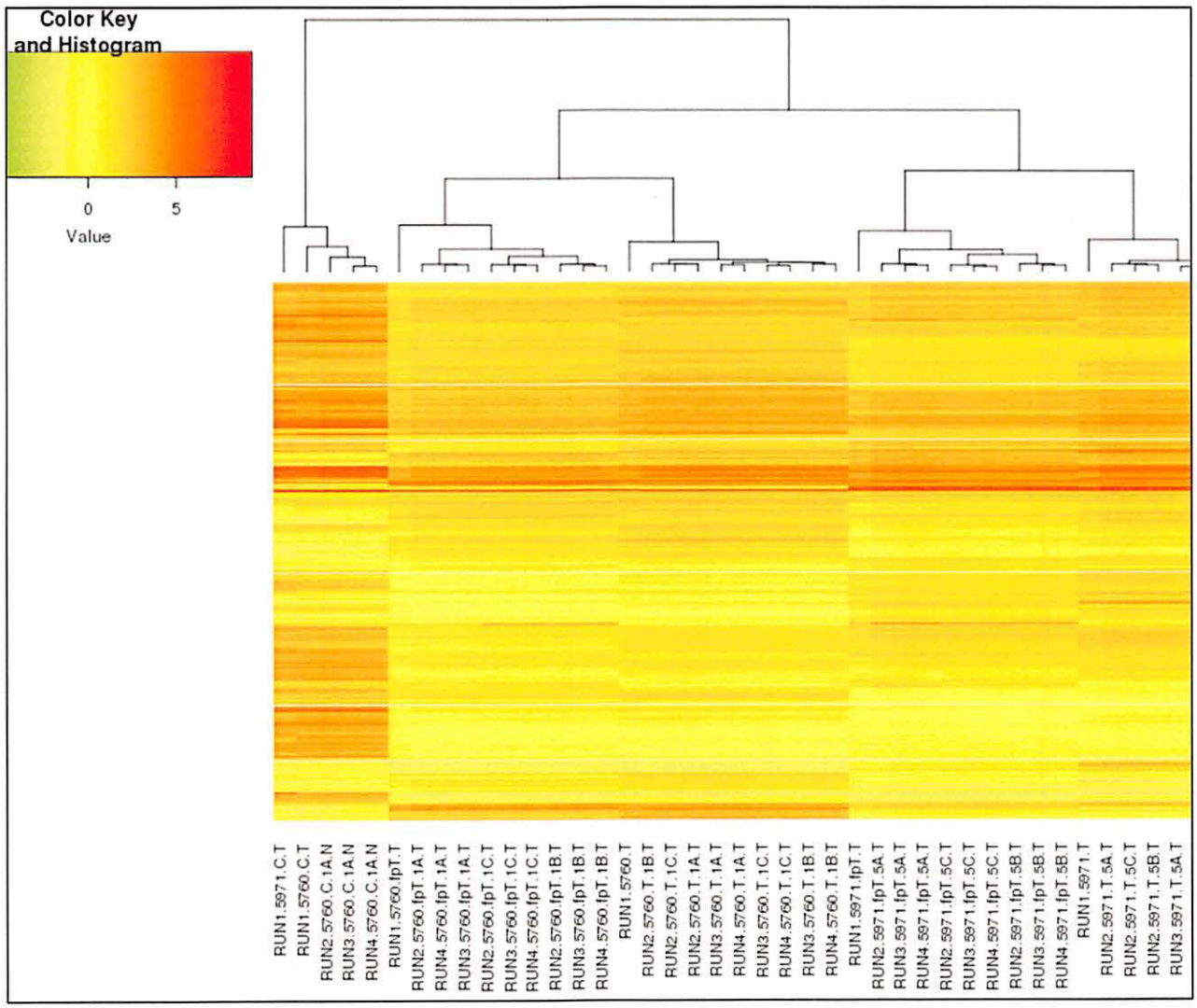


Table 5. Inter-run reproducibility of samples. Analytical metrics of samples sequenced at Run 1

MDA Sample ID	Tissue Type	RNA Conc. (ng/uL)	RNA Integrity Number (RIN)	RNA Quality (Ab_{260}/Ab_{280} ratio)	Total Reads	Number of Detected Genes	Percentage of Reads in Exonic Regions	Percentage of Reads in Intronic Regions	Percentage of Reads in Intergenic Regions
MDA-5760-T	Frozen	397.7	9.2	2.06	292.59	25758	0.841	0.089	0.069
MDA-5760-fpT	FFPE	382.0	2.4	2.05	292.88	25297	0.690	0.123	0.186
MDA-5760-C	PBMC	59.4	7.5	1.83	284.19	24678	0.770	0.138	0.091
MDA-5971-T	Frozen	668.6	8.3	2.09	326.62	26160	0.802	0.108	0.089
MDA-5971-fpT	FFPE	181.1	2.3	2.03	308.49	25703	0.702	0.127	0.170

Across FFPE vs FF reproducibility. Hierarchical clustering of the top 3000 differentially expressed genes between tumor samples and normal (PBMC) show FFPE and FF samples of MDA-5971 and MDA-5760 clustered together.

Figure 5B. Clustering Analyses of top 3000 differentially expressed genes in FF, FFPE tumors and PBMC.



The top 3000 differentially expressed genes from the previous sequencing run 1 clustered with differentially expressed genes from the current study (run 2, 3 and 4) as shown in **Figure 5B**. All RNA samples coming from PBMC clustered together. FF and FFPE samples of MDA-5760 clustered together from sequencing run at two different time points (run 1 at 1st-time point vs run 2, 3 and 4 at 2nd-time point). Similar observations were made for MDA-5971.

Step 2

5. **Pilot experiment:** A pilot test run of the samples was performed using 4 different lung cancer cases with FF and FFPE samples representing each case. The RNA extracted from each fresh-frozen

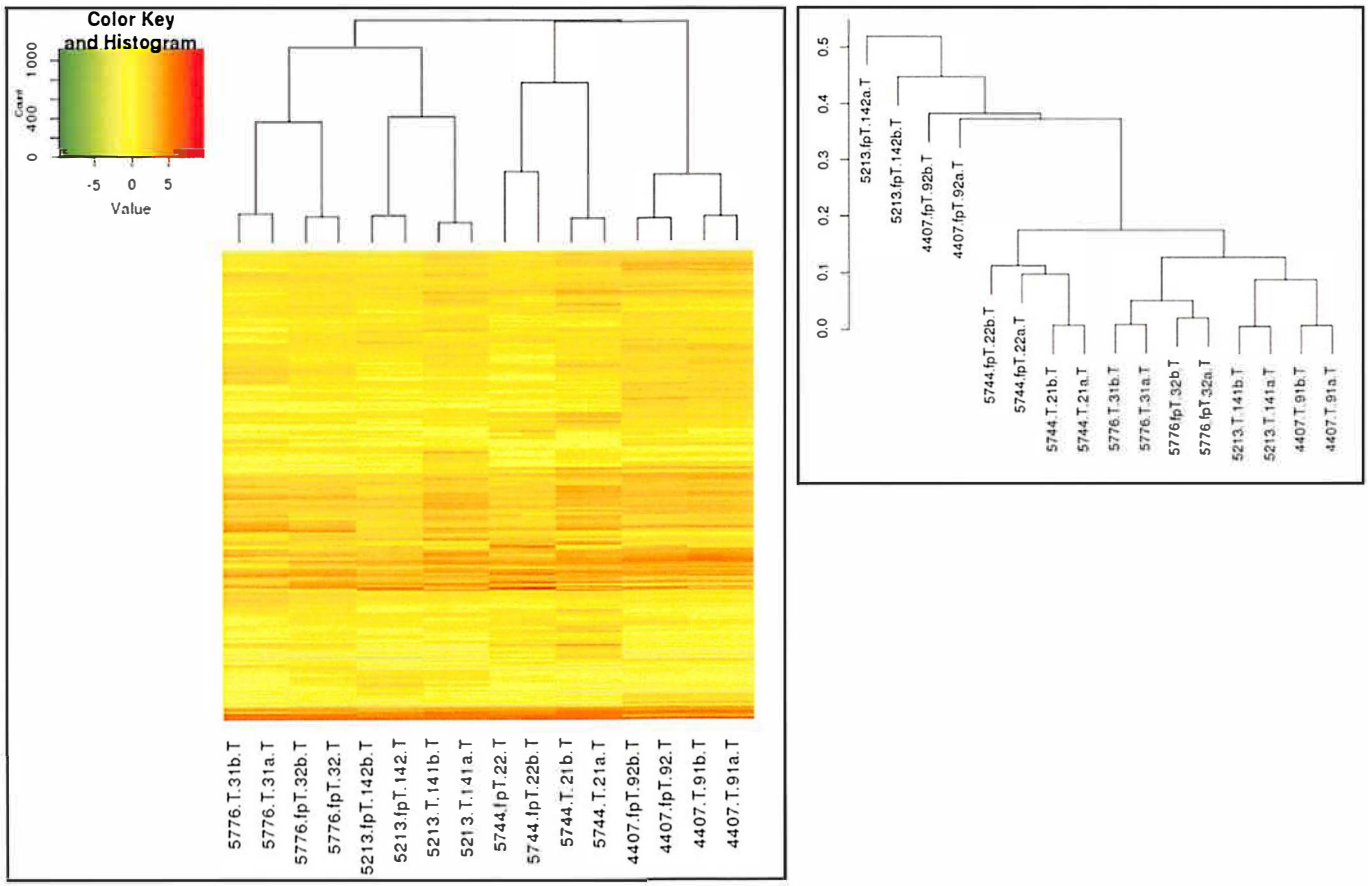
sample had DV200 values between 70-90%. The FFPE samples all had DV200 above 50. The RIN values were significantly low for the FFPE samples as shown in **Table 7**.

Table 7. Study Samples Information and Quality Control.

Sample ID	Tissue Type	RIN	DV200	Concentration (ng/ul)	Volume (ul)	Yield (ng)	260/280	260/230
MDA-5744-T	Frozen	8.3	87.95	5.71	22.0	125.6	2.09	2.11
MDA-5776-T	Frozen	1.0	72.10	7.41	22.0	162.9	2.07	2.05
MDA-4407-T	Frozen	6.2	72.27	6.81	24.0	163.4	2.10	1.84
MDA-5213-T	Frozen	7.4	90.27	6.70	24.0	160.8	2.11	2.11
MDA-5744-fpT	FFPE	1.9	53.12	12.92	15.5	200.0	1.86	2.00
MDA-5776-fpT	FFPE	1.4	62.2	16.54	10.0	165.4	1.81	2.00
MDA-4407-fpT	FFPE	1.2	52.16	14.47	10.0	144.7	1.81	2.19
MDA-5213-fpT	FFPE	1.5	55.21	19.11	10.0	191.1	1.82	2.16

Extracted RNA was aliquoted in duplicates and each set of aliquots was used to generate sequencing libraries 4 (tumor cases) x 2 (FFPE or FF) x 2 (duplicates) libraries. The sequenced data were analyzed as in the prior experiment. The top 3000 differentially expressed genes were obtained to determine concordance between FF and FFPE samples. **Figure 6** shows the clustering of genes in the FFPE and FF samples.

Figure 6. Clustering analyses of top 3000 differentially expressed genes in FF, FFPE tumors.



Conclusion: The analytical validation of RNA-Seq using FF and FFPE from lung tumors have been successfully performed. Data analysis using the set parameters show reproducible data that can be utilized for further interrogation.

References:

1. C. A. Dobin, D. F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras. 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15 – 21.

2. S. Anders, P. Theodor Pyl, W. Huber. 2014. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166 – 169.

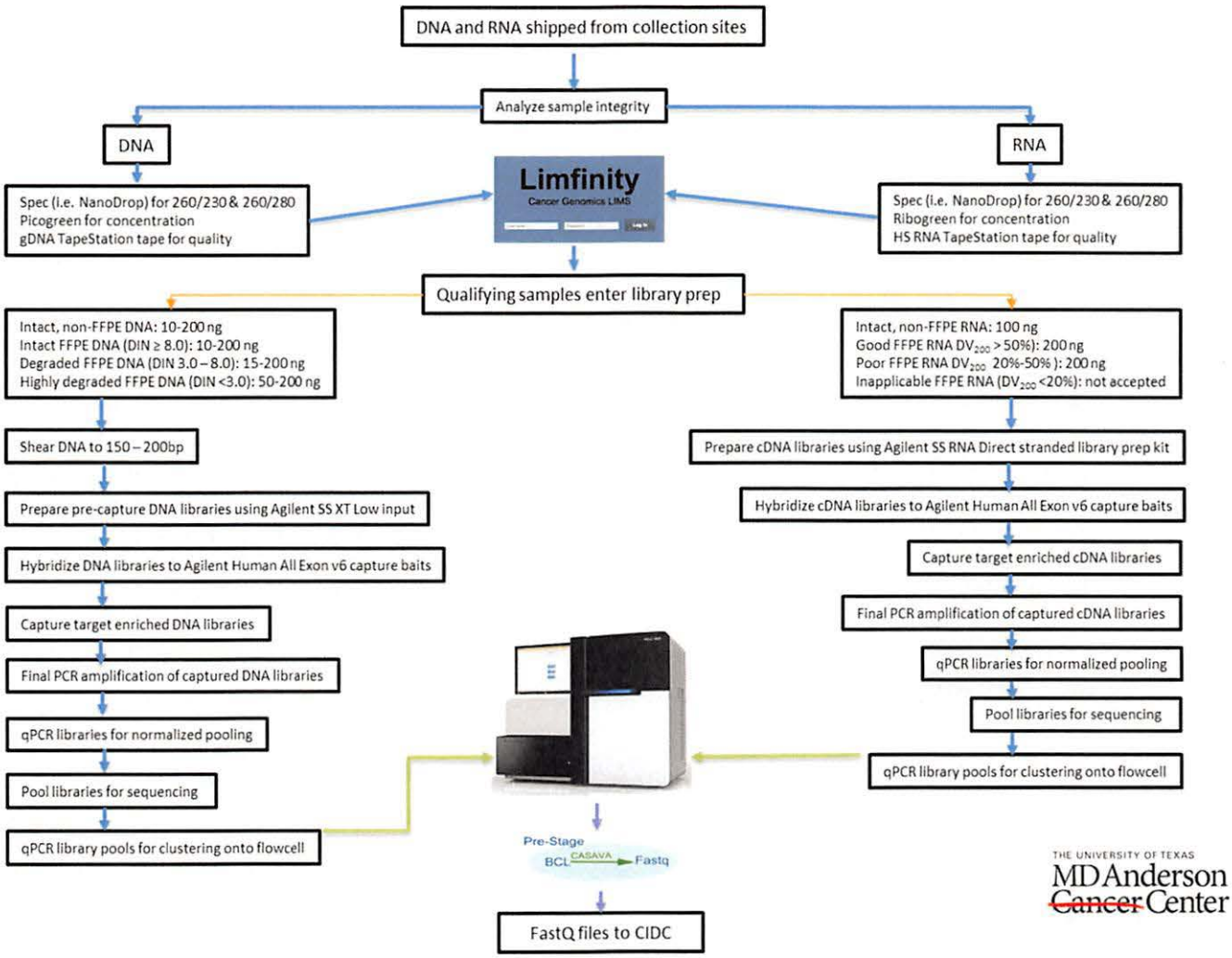
Director: Andy Futreal, BS, Ph.D., Professor and Chair;
Curtis Gumbs, Scientific Manager, Genomic Medicine
Jianhua (John) Zhang, Ph.D., Director, Computational Genomics, Genomic Medicine
Ignacio Ivan Wistuba, MD, Professor and Chair, Translational Molecular Pathology

Ignacio I. Wistuba, M.D.

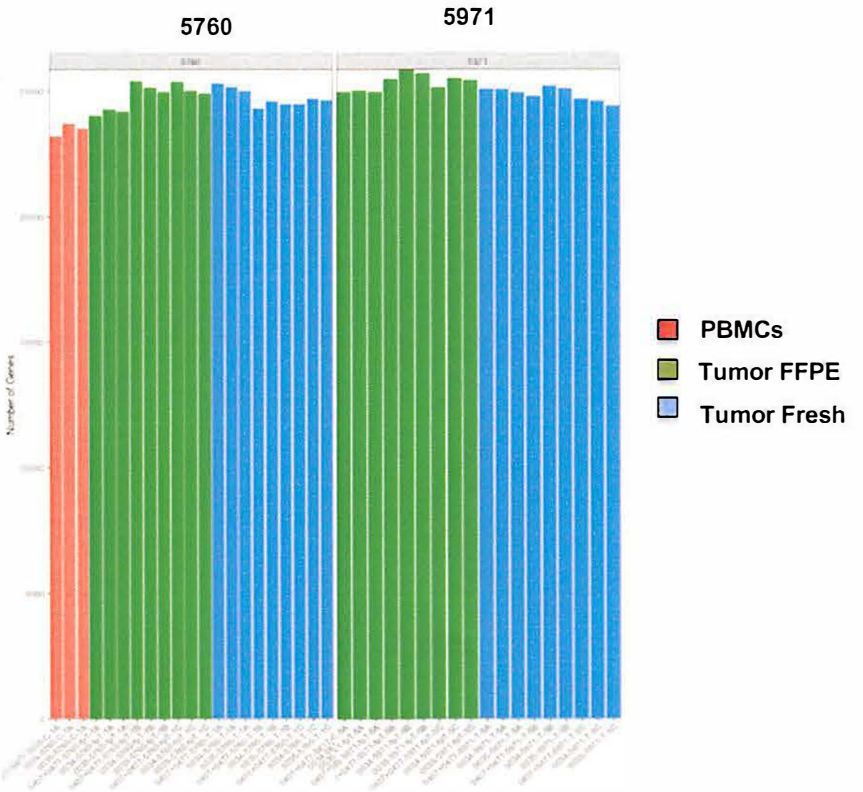
12/5/2019

Appendix

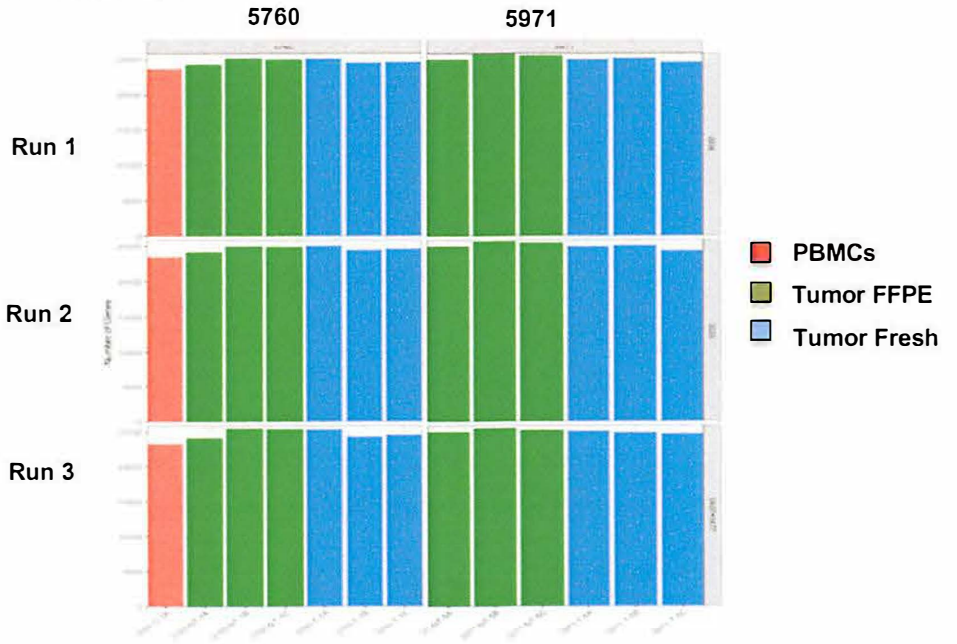
Supplemental Figure 1: MDACC RNA Seq workflow



Supplemental Figure 2A. Number of detected genes. (Each bar represents a patient and sample types are color-coded).



Supplemental Figure 2B. Number of detected genes. (Each bar represents a patient and sample types are color-coded).



Supplemental Figure 3. Percentage of reads in different regions: exonic, intronic, and intergenic.