invitrogen

Rapid Fluorescence Based Assessment of RNA Integrity Quality using Machine Learning

Scott T. Clarke[1], Kathy Free[1], Chris Vonnegut[1], Debra Gale[1], Dylan Poulsen[2]

[1]Cellular Labeling and Detection, Thermo Fisher Scientific, Eugene, Oregon, USA; [2] Washington College, Chestertown, Maryland, USA

ABSTRACT AND INTRODUCTION

Much attention has been focused on RNA transcript bias. The quality of isolated mRNA used for RNA-seq and other downstream analyses can contribute to sample specific bias of transcript abundance. If the library is prepared from polyA enrichment of degraded RNA or with reverse transcriptase using oligo dT priming then 3′ positional bias must be considered[1]

Isolated total RNA is routinely examined for its quality prior to committing to costly gene expression analysis. Since total RNA consists of ~85% rRNA, analysis primarily reflects rRNA quality and the mRNA quality is inferred. Typically, quality measurements consist of 260/280 nm ratio, 28S/18S rRNA ratio and capillary electrophoresis fragment analysis or gel-based separation and imaging methods. These methods which have become industry standards of quality to generate an RNA score such as RNA integrity number (RIN) between 0-10 with >8 being generally acceptable for RNA-seq. Generating RIN score requires experienced technical ability, specialized equipment, as well as additional cost and time prior to committing to the next step of building the library for sequencing.

In this presentation we examine the integrity of the RNA using a combination of specific fluorescent nucleic acid binding dyes which display differential binding specificity depending on the primary and secondary structure of the RNA. Presumably those structures change as the quality of RNA changes. For machine learning, a training data set of data was generated, followed by a test set of data used to determine accuracy. Data depicting the actual vs. predicted class are reported in a confusion matrix. Variable of importance analysis was used to determine the factors most important for predicting accurate RNA integrity quality (IQ).

The initial determination of accuracy of the machine learning based on test data, ranged between 92.7 - 95.7% over the entire range of RNA tested depending on the algorithm used. With further refinement to the machine learning, RNA IQ had an accuracy of ± 0.65 and ± 0.39 standard deviation (SD) for IQ range 0 – 10 using the test data set.

The accurate and low SD generated using nucleic acid binding dyes makes for a quick, simple and easy to use assessment of isolated RNA. The dye binding and fluorescent measurement is rapid and requires only a few minutes to generate. The dye/RNA fluorescence based determination of RNA quality was benchmarked against the industry standard of RIN score generated by capillary electrophoresis of the Bioanalyzer system (Agilent), Fragment Analyzer, and standard agarose gel based analysis to determine concordance and utility of this method.

Technical ability, cost and time investment are minimal allowing for rapid sample characterization prior to further analysis and commitment to downstream sequencing steps.

MATERIALS AND METHODS

Material and equipment for RNA quality and quantitation evaluation depicted in Figure 1. Bioanalyzer using RNA nanochip and Fragment Analyzer from Agilent (Santa Clara, CA). Oncomine™ Immune Response Research Reagent and Ion 530™ Chip from Thermo Fisher Scientific (Waltham, MA) . All other material and reagents from Thermo Fisher Scientific unless noted otherwise.

Figure 1. Methods used for assessing RNA quality

Figure **1: Methods for assessing RNA quality and quantitation 1a.** Qubit 4: fluorescence signals from nucleic acid dyes. **1b.** Fragment Analyzer: RQN score determined by capillary electrophoresis **1c.** Bioanalyzer: RIN score determined by capillary electrophoresis. **1d.** 2% agarose E-Gel EX electrophoresis **1e.** NanoDrop UV/vis absorbance

Figure 2. Predicted vs. actual RNA quality and quantity using surrogate data

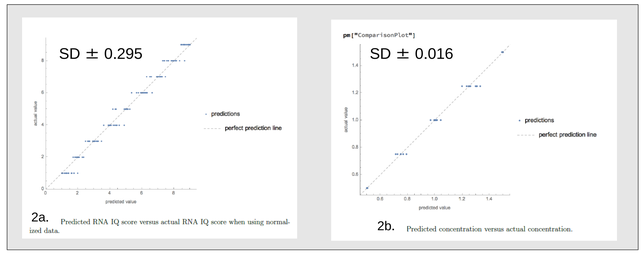
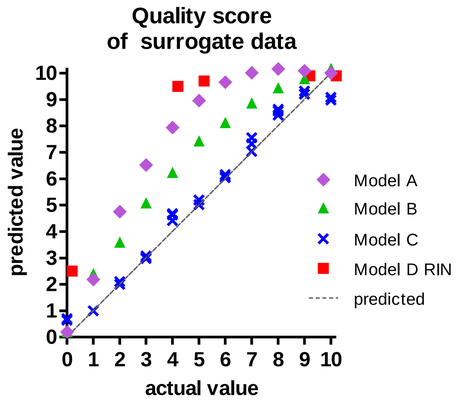
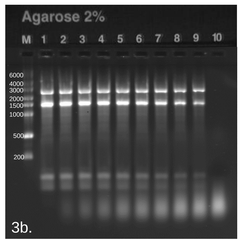


Figure 2: Correlation between actual and predicted RNA quality and quantity using surrogate data. Surrogate training data made from mixtures of intact and degraded RNA were used as input for machine learning. Test data made from similar mixtures were then used to determine the accuracy of prediction. The correlation between both quality and quantity measurements of the test values vs. predicted values were the highest observed with Model C. Standard deviation from the actual value is presented for both RNA quality (SD ± 0.295) (fig 2a.) and RNA quantitation (SD ± 0.016) (fig 2b).

Figure 3. Comparison of quality value of fluorescence based models and BioAnalyzer RIN score





lane of 2% gel	sample mixture	RNA Quality: Model B	RNA Quality: Model C machine learning	RNA Quality: Model D RIN	RNA Quantitation: Model C (ng/uL)	RNA Quantitation: BioAnalyzer quant (ng/uL)	RNA Quantitation: NanoDrop quant (ng/uL)
1	100%	10.18	9.05	9.9	118	136	109
2	90%	9.80	9.25	9.90	103	131	108
3	80%	9.45	8.54	N/D	88	112	108
4	70%	8.87	7.30	N/D	106	101	107
5	60%	8.13	6.09	N/D	102	84	107
6	50%	7.43	5.14	9.70	101	79	105
7	40%	6.24	4.57	9.50	104	60	105
8	30%	5.09	3.02	N/D	109	49	107
9	20%	3.60	2.03	N/D	107	40	107
	10%	2.41	1.00	N/D	105	35	103
10	0%	0.21	0.66	2.50	111	26	103
	blank	10.46	5.21	N/D	65	10	1

3c.

Quality ranking vs. predicted value
0-5%
5-10%
>10%

Quantitation ranking vs. NanoDrop
0-10%
10-20%
>20%

Figure 3. Various analytical methods used for evaluating quality and quantity of surrogate data generated from mixtures of RNA. Graph depicts RNA quality compared to perfect prediction value (fig. 3a). 2% agarose gel showing sample mixtures used for training and test data (fig. 3b) was analyzed for quality using mathematical Models A and B, and machine learning Model C comparing to Model D, BioAnalyzer RIN score. Theoretical percentage was used as reference to evaluate the closeness of fit for quality. Quantitation was determined in Model C machine learning and Model D (BioAnalyzer). Nanodrop was used as reference standard to evaluate closeness of fit for quantitation. Table is color coded based on percentage from expected value (fig. 3c). BioAnalyzer does not accurately evaluate mixtures of RNA. Model C machine learning shows higher correlation between actual and predicted values than seen with other models. The deviation from the true value is largest at the points of highest and lowest % mixtures of RNA.

Figure 4. Distribution of training data and deviation from prediction

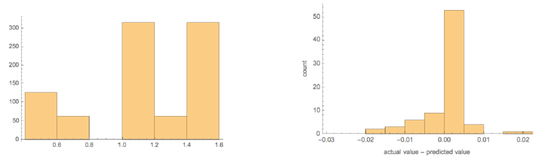
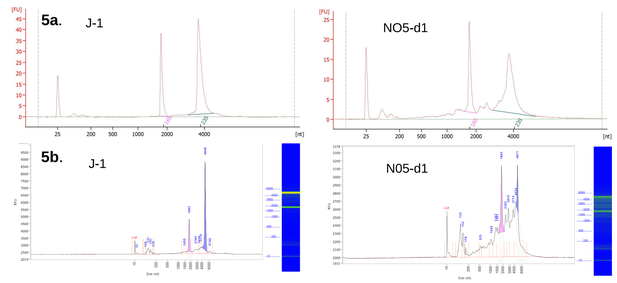


Figure 4: Number of occurrences of each concentration in the training data set. Correlation of predicted vs. actual value can be improved by increasing the occurrence of training data samples (**4a**). Histogram of deviation between actual and predicted concentration (**4b**).

Figure 5. Characterizing and sequencing of intact and partially degraded RNA



well	sample	RNA Quality: Model B quality	RNA Quality: Model C machine quality	RNA Quality: Model D RIN	RNA Quality: Fragment Analyzer (RQN)	RNA quantitation: Model C machine quant (ng/uL)	RNA quantitation: Model D BioAnalyzer quant (ng/uL)	RNA quantitation: Fragment Analyzer quant (ng/uL)	RNA quantitation: NanoDrop quant (ng/uL)	RNA quantitation: Supplier quant (ng/uL)
1	iXS-1	9.3	6.6	10	10	105	144	138	140	
2	ji-1	7.9	7.4	10	10	71	90	94	79	
3	NO5-d1	8.5	8.2	9.1	7.6	111	104	112	106	
4	NO5-1	8.7	8.2	9.2	6.9	101	107	89	114	
5	C1	8.5	5.9	2.5	3.2	114	98.2	65	138	89.6
6	D1	7.8	5.8	1.9	1.5	106	109.3	80	100	74.4
7	S1	7.5	6.3	1.9	1.1	103	48	65	84	65.2
8	E5-1	10.2	5.1	N.D.	N/A	70	3	N/A	3	N.D.
9	Std 3	10	N/A	N/A	N/A	N/A	187	N/A	105	
10	Std 2	0	N/A	2.3	N/A	N/A	21	N/A	100	

5d.

Quality ranking compared to RQN
0-5%
5-10%
>10%

Quant. ranking compared to NanoDrop
0-10%
10-20%
>20%

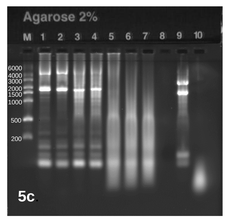


Figure 5: Characterizing RNA prior to downstream applications: Two samples of total RNA derived from Jurkat cells were characterized and analyzed for differences in gene expression of the 398 genes in the Oncomine™ Immune Response Research Assay. One sample (J-1) was stored at -80˚C and remained intact while the other sample (NO5-d1) was stored for an extended period at -20˚C and had partially degraded. NanoDrop using UV absorbance for quantitation together with BioAnalyzer (5a), Fragment Analyzer (5b), and the other models using fluorescence signal (table 5d) along with gel electrophoresis (5c) were used to characterize the material used to prepare the sequencing library. Correlation between models for RNA quality score is poor. Correlation between RNA quantity is closer for some of the models.

Figure 6. preparation of sequencing library

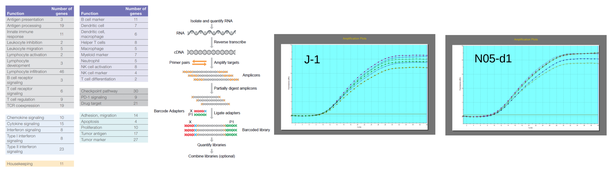
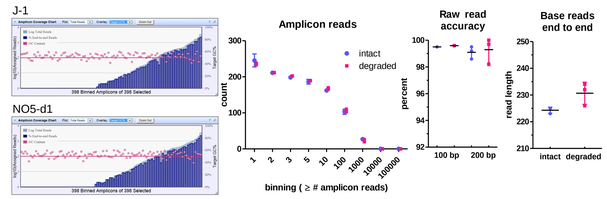


Figure 6: Preparing and quantitating the library. Prior to running the targeted gene expression assay on the Ion™ next-generation sequencing (NGS) platform the library is quantitated using Ion Library Taqman quantification kit. The pan-cancer gene expression assay targets 398 genes relevant to the tumor microenvironment.

Figure 7. Analysis of Oncomine Immune Respond Research Assay panel



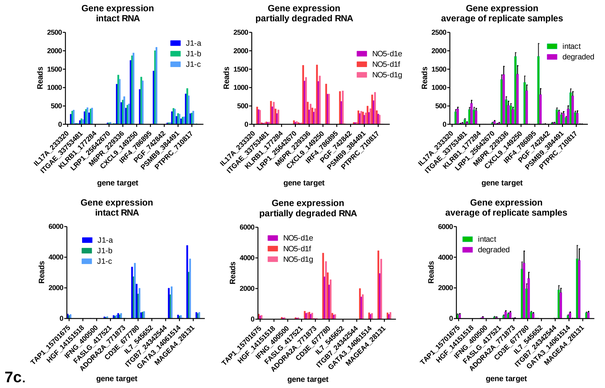


Figure 7: Oncomine™ Immune Response Research Assay: Amplicon coverage chart for intact and degraded RNA (7a). Binned number of reads and Raw Read Accuracy are equivalent for the two samples of RNA (7b). Two subsets of gene expression panel showing equivalent levels between intact RNA (J-1) and partially degraded RNA (NO5-d1). No statistical difference between replicates and samples. The level of degradation of NO5-d1 was not sufficient to alter gene expression values (7c).

CONCLUSIONS

- Machine learning can accurately predict RNA IQ score and quantitation of test data when based on surrogate training data used to develop the algorithm.
- Differential binding of RNA binding dyes results in fluorescence signals that reveal RNA quality over the range of the training data used.
- For training data to better predict RNA quality the input must simulate the changes that occur in secondary structure that occurs when RNA is degraded.
- The degraded RNA tested did not have significant alteration in gene expression over intact RNA.
- A wider range of RNA degradation will be needed to observe a correlation between RNA quality and an impact on transcript abundance in RNA sequencing.

REFERENCES

1. Love, M.I., Hogenesch, J. B., Irizarry, R.A. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* 1-15. doi:10.1038/nbt.3682

ACKNOWLEDGEMENTS

The authors would like to thank Sandra Lui, Maryam Shenasa and Janice Au-Young of the Clinical Next Gen Sequencing Division of Thermo Fisher Scientific for their expert advice and for sequencing and analyzing the library.

ThermoFisher SCIENTIFIC