

Mixed effects model for assessing RNA degradation in Affymetrix GeneChip experiments

Kellie J. Archer, Ph.D.

Suresh E. Joel

Viswanathan Ramakrishnan, Ph.D.

Department of Biostatistics

Virginia Commonwealth University

e-mail: kjarcher@vcu.edu

Assessing Quality

- 2D Spatial images
- Boxplots
- MA Plots
- RNA Purity
 - A_{260}/A_{280} ratio (protein contaminants)
 - A_{260}/A_{270} ratio (phenol/organic contaminants)
 - ribosomal RNAs
- RNA Integrity
 - 3':5' ratios
 - *28S/18S* ratio (electropherogram)
- Linearity

Assessing quality from a hybridized microarray

- Detection call for ribosomal RNAs (rRNAs) *28S* and *18S*
- Linearity
- 3':5' ratios

Transcription

- Process of *in vitro* transcription begins with reverse transcriptase starting the cDNA synthesis from an oligo(dT) primer that anneals to the 3' end the mRNA template molecule and continuing toward the 5' end of the template transcript.
- Transcription may not continue to completion.
- 3':5' ratio assesses the degree to which genes were transcribed.

3':5' Ratios

- Affymetrix GeneChips include probe sets whose primary purpose is to assess the quality of transcription.
- For example, the HG-U133A and HG-Focus GeneChips include probe sets which interrogate both the 3' and 5' end of the same gene (*GAPDH*, *β-actin*, *ISGF*, *18S*, and *28S*).

3':5' Ratios

- Problem during RNA extraction where the starting RNA was not of full length;
- Problem during cDNA synthesis reaction where mRNA may not have been fully converted to cDNA;
- Problem during IVT/Biotin labeling reaction where cDNA was not properly converted to biotinylated cRNA.

	Criteria for 3':5' ratio	Source
Affymetrix	< 3	http://www.affymetrix.com/support/technical/manual/expression_manual.affx
UCLA DNA Microarray Core Facility	< 2	http://www.genetics.ucla.edu/microarray/EDCGuidelines_for_checking_the_.htm
W.M. Keck Foundation Biotechnology Resource Laboratory	< 3	http://info.med.yale.edu/wmkeck/affymetrix/analysis.htm
University of Michigan's Microarray Core Facility	< 3	http://www.umich.edu/~caparray/GeneChip.html

Expression Quantification

- Affymetrix Genechip is an oligonucleotide array consisting of a several perfect match (PM) and their corresponding mismatch (MM) probes that interrogate for a single gene.
 - PM is the exact complementary sequence of the target genetic sequence, composed of 25 base pairs
 - MM probe, which has the same sequence with exception that the middle base (13th) position has been reversed
 - There are roughly 11-20 PM/MM probe pairs that interrogate for each gene, called a probe set

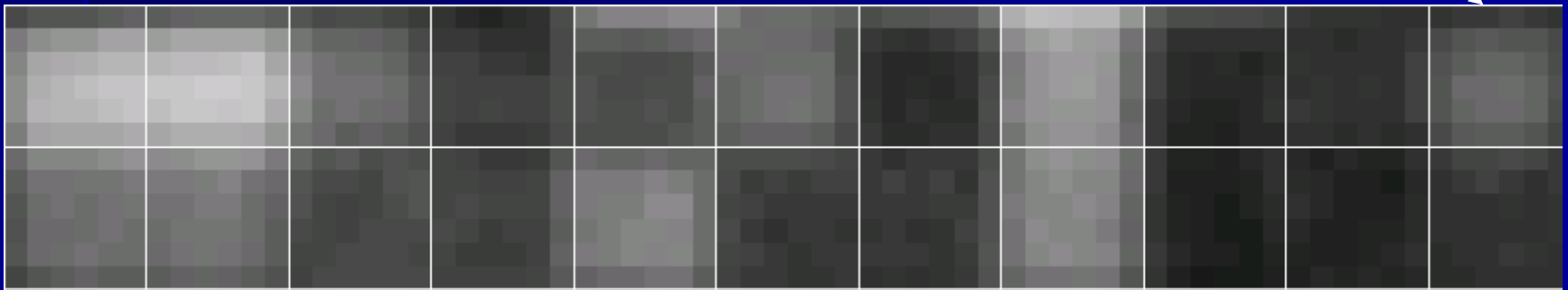
Expression Quantification

11 – 20 Probe Pairs interrogate each gene

PM and MM intensities are combined to form an expression measure for the probe set (gene)

GCGCCGGCTGCAGGAGCAGGAGGAG

PM



GCGCCGGCTGCACGAGCAGGAGGAG

MM

Expression Quantification

- Average difference
- Model Based Expression Index (Li & Wong, 2001)
- Robust Multiarray Average (Irizarry et al, 2003)
- MAS 5.0 method (Hubbell et al, 2002)

Illustrative Data

- Previously published data collected at the University of Tübingen to assess the impact of RNA degradation on microarray gene expression in renal cell carcinoma patients (Schoor et al, 2003) will be used to illustrate RNA quality assessment.

Illustrative Data

- RNA samples of tumor and normal tissue at freshly isolated and two different degradation states were used.
- This study included 9 GeneChips
 - 4 HG-U133A GeneChips
 - 5 HG-Focus GeneChips

Illustrative Data

	Source	Degradation state	GeneChip
TA-U	tumor	A	U133A
NA-U	normal	A	U133A
TA-F1	tumor	A	Focus
TA-F2	tumor	A	Focus
NA-F	normal	A	Focus
TB-F	tumor	B	Focus
NB-F	normal	B	Focus
TD-U	tumor	D	U133A
ND-U	normal	D	U133A

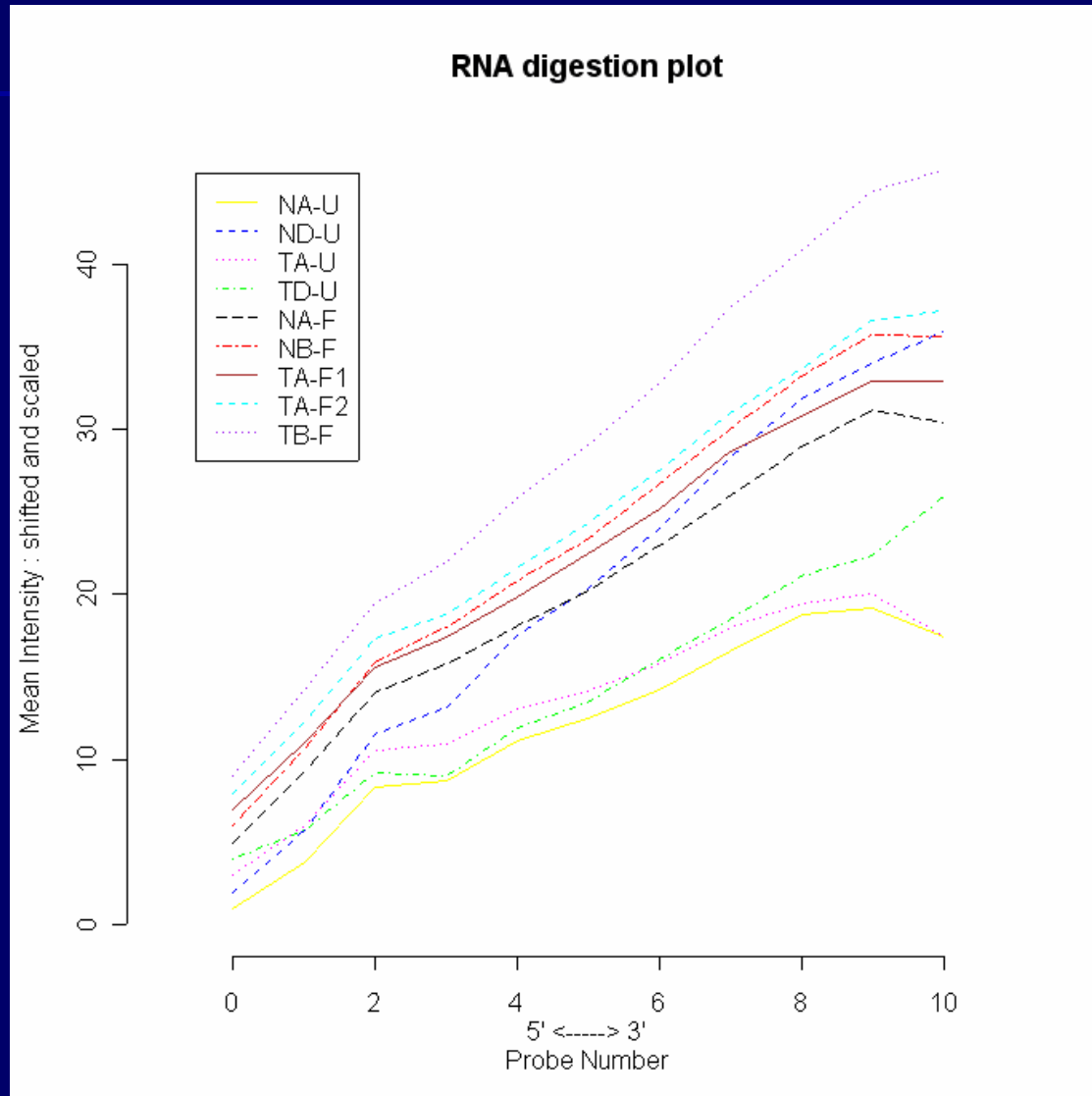
3':5' Ratios

GeneChip	MAS 5.0 3':5' ratio		RMA 3':5' ratio	
	<i>GAPDH</i>	<i>β- actin</i>	<i>GAPDH</i>	<i>β- actin</i>
NA-U	1.07	1.13	1.07	0.98
TA-U	0.91	1.08	1.04	0.94
NA-F	1.91	6.04	1.09	1.26
TA-F1	1.34	3.87	1.05	1.15
TA-F2	1.17	5.44	1.05	1.19
NB-F	2.79	9.04	1.13	1.31
TB-F	3.52	22.82	1.15	1.41
ND-U	6.94	13.58	1.35	1.46
TD-U	9.05	11.46	1.37	1.51

RNA digestion plot

- A recently proposed method for assessing sample degradation is the RNA digestion plot (Gautier et al, 2004).
- Utilizes chip design: Probes interrogating a transcript are arranged in order of their interrogation position.
- Plot of overall mean expression by probe interrogation position (`_at1's`, `_at2's`, ..., `_at11's`).

RNA digestion plot

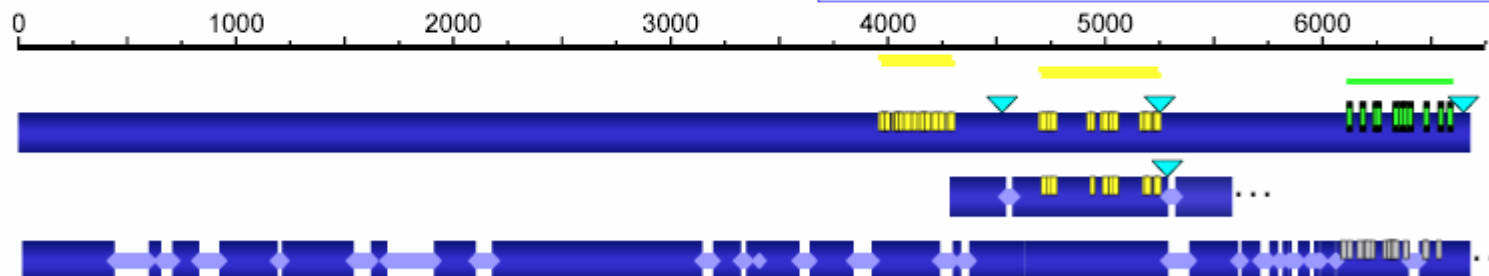




HG-U133:200597_at

Gene: EIF3S10

eukaryotic translation initiation factor 3, subunit 10 theta,
150/170kDa



200596_s_at

200595_s_at

200597_at

Computer representation of images

- Consider the computer image to be a two dimensional array (matrix) of numbers.
- The smallest element of the image is a pixel.
- For an image with $M \times N$ pixels, each pixel has location (x,y) .
- Each pixel has an intensity value $f(x,y)$, and the size of the pixel is $\Delta x * \Delta y$.

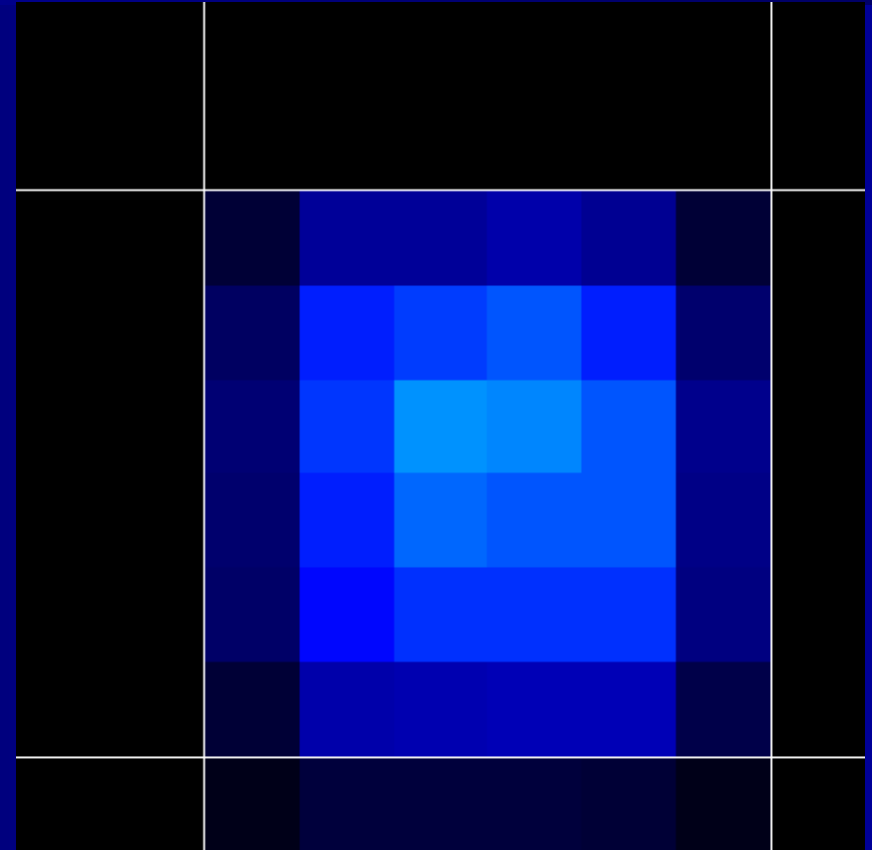
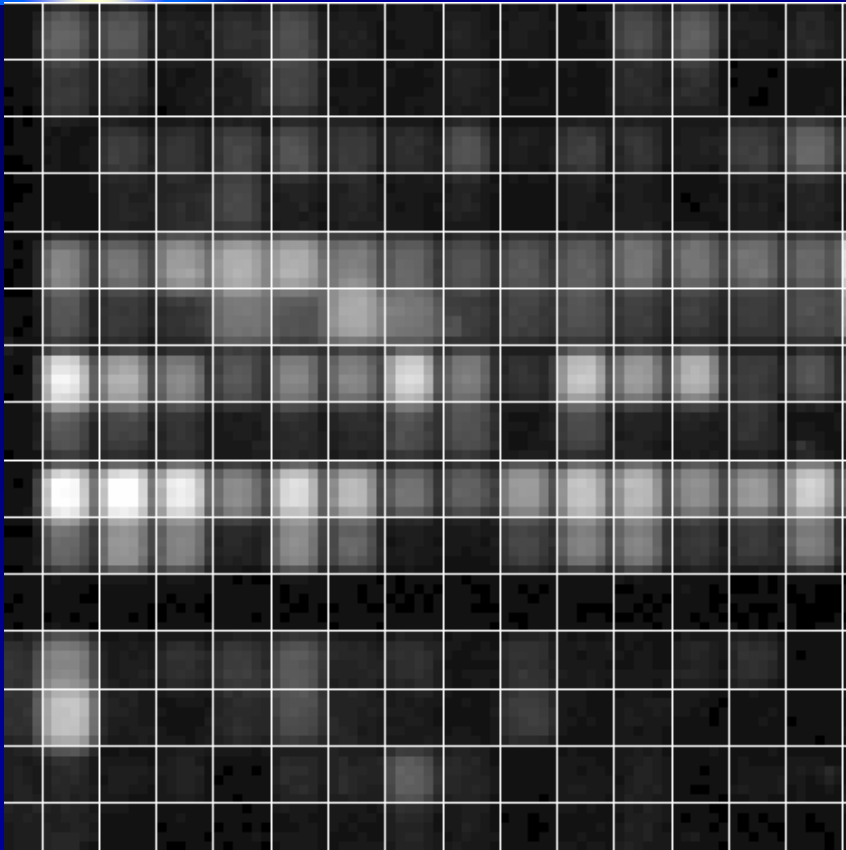
Computer representation of images

- For a monochromatic image, $f(x,y)$ is an integer called a gray value where
$$f = \{f(x,y): x = 0, 1, \dots, M-1; y=0, 1, \dots, N-1.\}$$
- Therefore, each $f(x,y)$ represents the brightness of a small picture element, called pixel, at location (x,y) .

Image Analysis

- Addressing, Segmentation, Intensity extraction
- After identifying the location, size and shape of each probe and identifying background versus foreground pixels, gene expression values are calculated using some function (e.g., mean, median, 75th percentile, etc.) of the observed pixel values within each segmented area.

Image Analysis: Pixel Level Data



6 x 6 matrix of pixels for each PM and MM probe
HG-U133A GeneChip

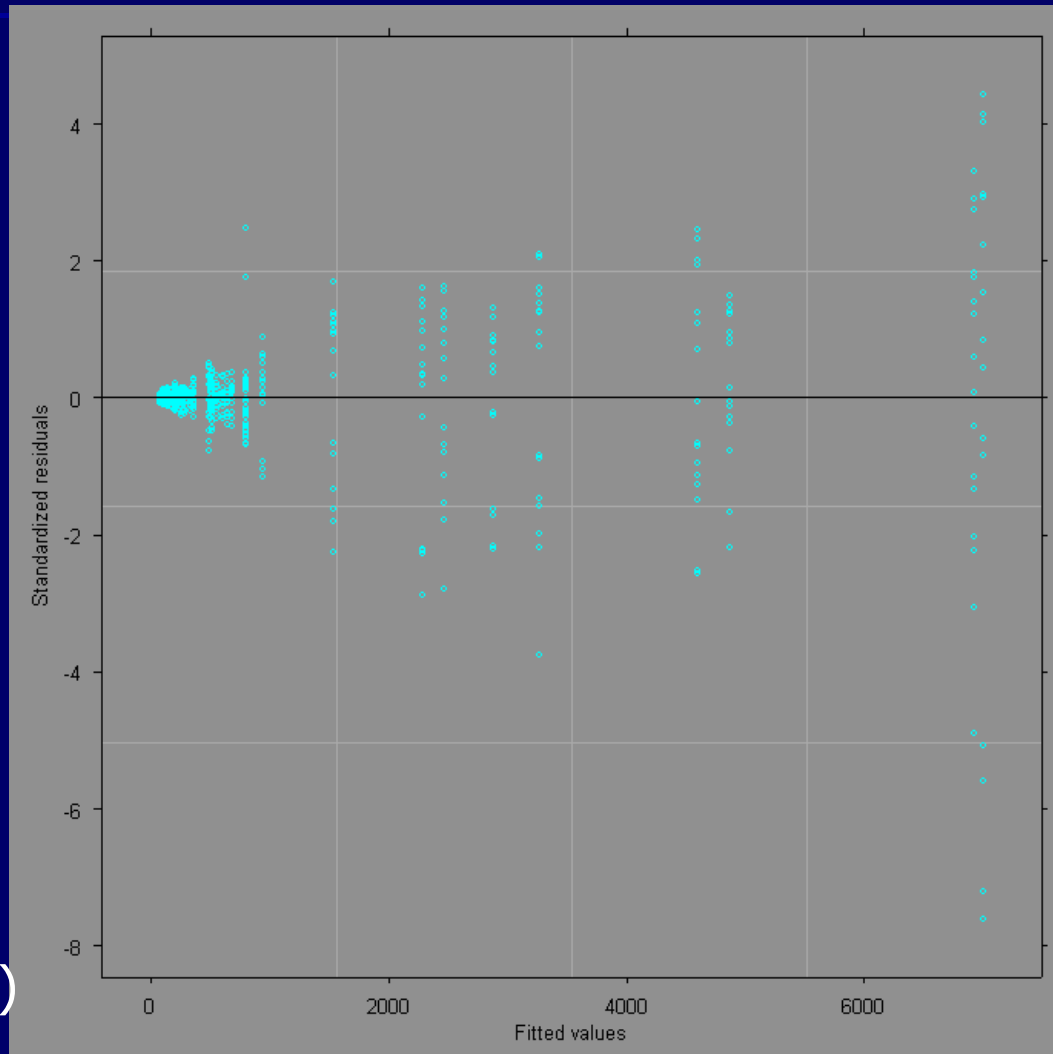
Mixed Effects Model for Assessing 3':5' Ratios

- Inherent hierarchical structure to GeneChip data
 - pixels are nested within probes
 - probes are nested within probe sets
- Assess RNA degradation by fitting a mixed effects model
 - probe set is fixed effect of interest
 - PM probe is random effect
 - pixels are a subsample nested within PM probe

Mixed Effects Model for Assessing 3':5' Ratios

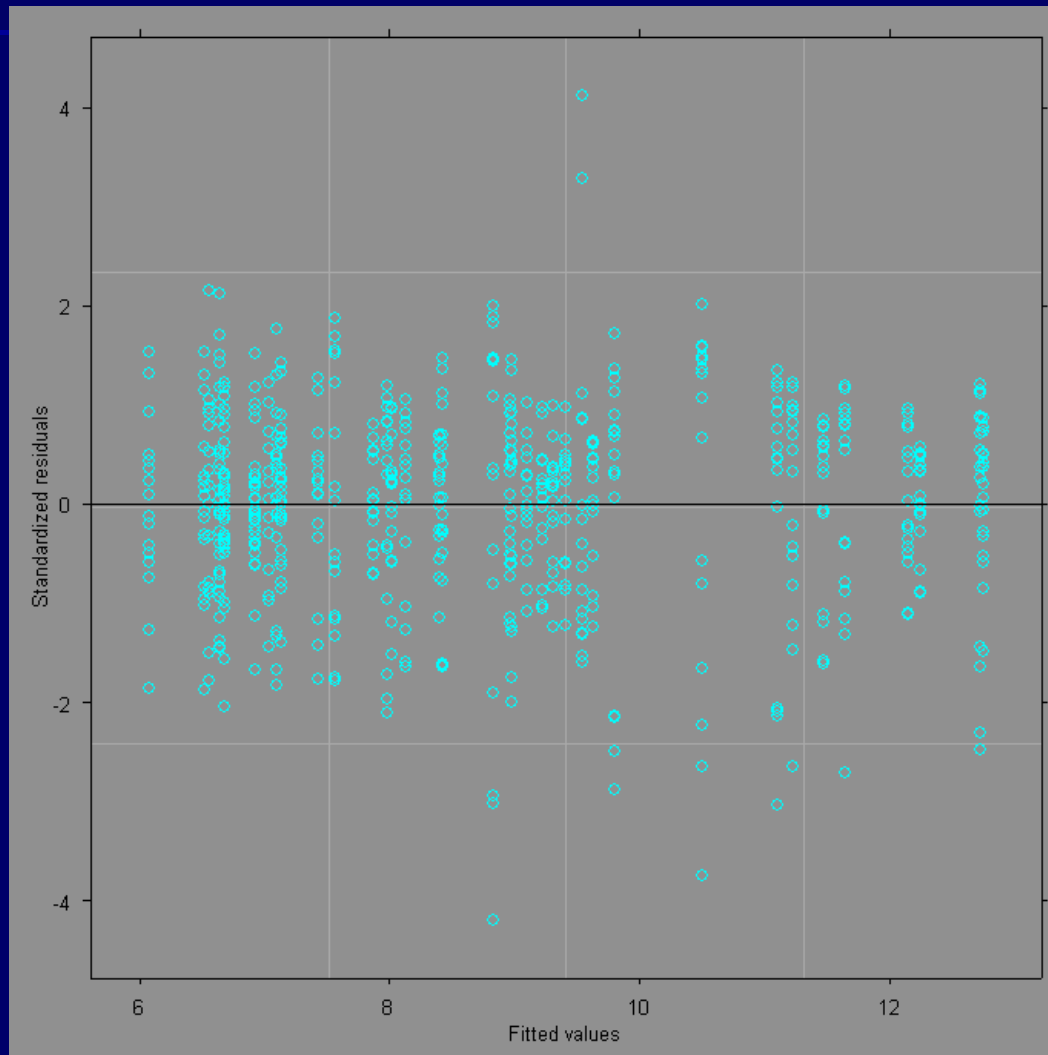
- Model: $y_{ij} = \mu + \beta_i + b_{i(j)} + \varepsilon_{ij}$
 - where y_{ijk} represents the signal intensity
 - μ represents the overall mean
 - β_i represents the fixed probe set effect
 - $b_{i(j)}$ represents the random PM probe effects assuming $b_i \sim N(0, \sigma_b^2)$ for $i=1, \dots, 20$
 - ε_{ijk} represents the error assuming $\varepsilon_{ijk} \sim N(0, \sigma^2)$

Model assessment: absolute signal intensities



(ND-U GeneChip)

Model assessment: \log_2 signal intensities



(ND-U GeneChip)

Mixed Effects Model for Assessing 3':5' Ratios

- The antilog of the resulting contrast and associated 95% confidence interval for the fixed effect parameter estimates are reported.
- This provides an interpretation on the original scale (3':5' Ratio).
- p-value from test of $H_0: 3':5' \text{ Ratio} = 1$ is equivalent to the p-value from the test $H_0: \log \text{ difference} = 0$.

Estimates from mixed effect models

	<i>β-actin</i>			<i>GAPDH</i>		
Chip	3':5'	95% CI	p-value	3':5'	95% CI	p-value
NA-U	0.80	(0.38,1.69)	0.56	1.66	(0.86,3.22)	0.13
TA-U	0.60	(0.31,1.17)	0.13	1.38	(0.16,11.55)	0.76
NA-F	3.18	(1.36,7.42)	0.009	2.43	(1.29,4.54)	0.007
TA-F1	2.49	(1.17,5.31)	0.02	1.65	(0.89,3.06)	0.11
TA-F2	2.97	(1.33,6.64)	0.009	1.78	(1.00,3.18)	0.05
NB-F	4.91	(2.21,10.90)	<0.001	3.46	(1.85,6.46)	<0.001
TB-F	8.92	(4.10,19.38)	<0.001	4.43	(2.31,8.48)	<0.001
ND-U	3.12	(1.41,6.89)	0.006	7.57	(4.35,13.19)	<0.001
TD-U	2.81	(1.42,5.55)	0.004	6.45	(3.87,10.74)	<0.001

Conclusions

- 3':5' ratio depends on probe set expression summary method used.
- RNA digestion plots are somewhat subjective in their interpretation.
- The proposed mixed effects models using PM pixel level intensities provides framework for estimating the 3':5' ratios and associated confidence interval and p-value.

Acknowledgements

- Catherine I. Dumur, Ph.D.
- Oliver Schoor, University of Tübingen
- DevNet support at Affymetrix
- R development team
- Bioconductor development team, affy library
- Pinheiro and Bates, n1me library

References

- Affymetrix, Santa Clara, CA. (2003) "Genechip expression analysis technical manual." Retrieved 05/17/04.
(http://www.affymetrix.com/support/technical/manual/expression_manual.affx).
- Castro-Vargas, E., UCLA DNA Microarray Core Facility. (2001) "Guidelines for checking the quality of your genechip array." Retrieved 05/17/04.
([http://www.genetics.ucla.edu/microarray/EDCGuidelines for checking the .htm](http://www.genetics.ucla.edu/microarray/EDCGuidelines_for_checking_the_.htm)).
- W.M. Keck Foundation Biotechnology Resource Laboratory, New Haven, CT. (2003) "Genechip expression analysis experiments." Retrieved 05/17/04.
(<http://info.med.yale.edu/wmkeck/affymetrix/analysis.htm>).
- University of Michigan Microarray Core Facility, Ann Arbor, MI. (2004) "Genechip expression analysis experiments." Retrieved 05/17/04.
(<http://www.umich.edu/~caparray/GeneChip.html>).
- Li, C., Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of the National Academy of Science*, **98**: 31-36.
- Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P. (2003) Summaries of affymetrix genechip probe level data., *Nucleic Acids Research*, **31**: e15.
- Hubbell, E., Liu, W.-M., Mei, R. (2002) Robust estimators for expression analysis, *Bioinformatics*, **18**: 1585-1592.

References

- Schoor, O., Weinschenk, T., Hennenlotter, J., Corvin, S., Stenzel, A., Rammansee, H.-G., Stevanovic, S. (2003) Moderate degradation does not preclude microarray analysis of small amounts of rna., *BioTechniques*, **35**: 1192-1201.
- Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. (2004) Affy-analysis of affymetrix genechip data at the probe level, *Bioinformatics*, **20**: 307-315.
- Liu, W.-M., Mei, R., Di, X., Ryder, T.B., Hubbell, E., Dee, S., Webster, T.A., Harrington, C.A., Ho, M.H., Baid, J., Smeekens, S.P. (2002) Analysis of high density expression microarrays with signed-rank call algorithms, *Bioinformatics*, **18**: 1593-1599.
- Pinheiro, J.C., Bates, D.M. (2000) *Mixed-effects models in s and s-plus*. Springer-Verlag, New York.
- Ihake, R., Gentleman, R. (1996) R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, **5**: 299-314.