BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 1 2013, pages 129–131
doi:10.1093/bioinformatics/bts629

Gene expression Advance Access publication October 24, 2012

AffyRNADegradation: control and correction of RNA quality effects in GeneChip expression data

Mario Fasold[1,2,*] and Hans Binder[1,2]

[1]Interdisciplinary Center for Bioinformatics, Universität Leipzig, D-4107 Leipzig, Haertelstr. 16-18, Germany and [2]Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Gene expression experiments aim to accurately quantify thousands of transcripts in parallel. Factors posterior to RNA extraction can, however, impair their accurate representation. RNA degradation and differences in the efficiency of amplification affect raw intensity measurements using Affymetrix expression arrays. The positional intensity decay of specifically hybridized probes along the transcript they intend to interrogate is used to estimate the RNA quality in a sample and to correct probe intensities for the degradation bias. This functionality, for which no previous software solution is available, is implemented in the R/Bioconductor package `AffyRNADegradation` presented here.

Availability: The package is available via Bioconductor at the URL http://bioconductor.org/packages/release/bioc/html/AffyRNADegradation.html

Contact: Fasold@izbi.uni-Leipzig.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 27, 2012; revised on September 26, 2012; accepted on October 16, 2012

1 INTRODUCTION

A basic assumption in gene expression experiments is that the obtained data represent a snapshot of transcript abundances within the original sample. However, several effects can distort the amount of RNA during sample extraction and preparation, and thereby impair the reliability of those measurements. RNases introduced by improper purification or incautious sample handling can degrade the rather unstable RNAs during storage (Fleige and Pfaffl, 2006). Also, the amplification of RNA mandatory to most RNA analytics differs in its efficiency and can therefore lead to variation in transcript yield and lengths (Ma *et al.*, 2006).

Gene expression experiments are frequently conducted using high-density microarrays. Because of the importance of RNA quality for the reliability of the results, it is advised to check the integrity of the RNA before hybridization to the array. RNA integrity (RIN) values (Schroeder *et al.*, 2006) that are determined on the basis of an electropherogram trace have become the standard measure of RNA quality. Samples with RIN values >7 should be discarded.

Researchers increasingly conduct large-scale meta-analysis on the plethora of publicly available microarray data. For these data, RNA quality measures are mostly not available. However, it is strongly advised to identify and to remove low RNA-quality experiments, as they can lead to erroneous results. Methods to estimate RNA quality directly from microarray data are thus required. Existing options are the use of 3′/5′ intensity ratios of control probe sets included on the microarray, as well as 3′/5′-summary degradation measures as provided by software tools such as the `affy` package (Gautier *et al.*, 2004). Both methods have been shown to have drawbacks under circumstances that are relevant in large-scale studies (Fasold and Binder, 2012). Particularly, 3′/5′ control probes might be affected by saturation, whereas affyslope estimates are affected by background hybridization. Both methods are prone to systematically overestimating RNA quality.

Beyond strict quality control and the removal of bad-quality samples, the continuous levels of RNA quality transform into a gray area of biased expression results with questionable reliability. It has been previously found that, although moderate levels of RNA degradation are tolerated by differential expression analysis, especially long targets provide erroneous results.

In this work, we present an R package that assesses RNA quality of Affymetrix expression data. It provides a RNA quality measure that overcomes the drawbacks of existing methods by strictly referring to specific hybridization. Furthermore, it enables correction of the 3′ probe intensity bias for improved downstream analysis.

For illustration, we here use data from an experiment done by Archer *et al.* (2006) where the same cell extract has been used for multiple microarray hybridizations, however, either prepared with RNeasy to remove RNA degrading enzymes, or not.

2 FUNCTIONALITY

On Affymetrix 3′, expression arrays up to 16 probes of length 25 nt interrogate each transcript. Most of these probes cover a specific region located within 600 nt distance to the 3′ end of the transcripts. RNA samples are usually prepared using an *in vitro* transcription labeling and amplification assay with primers starting at the 3′ poly-A tail of the source mRNA. Both degradation of mRNA as well as effectiveness of the amplification assay are thus captured by multiple probe measurements for each transcript.

2.1 Analyzing RNA degradation and amplification

Limited RNA quality of a given sample leads to intensity differences between probes located at the 3′ end and those located closer toward the 5′ end of the mRNA. The so-called

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Universitaet Leipzig, Institut fuer Informatik/URZ, Bibliothek on January 29, 2013

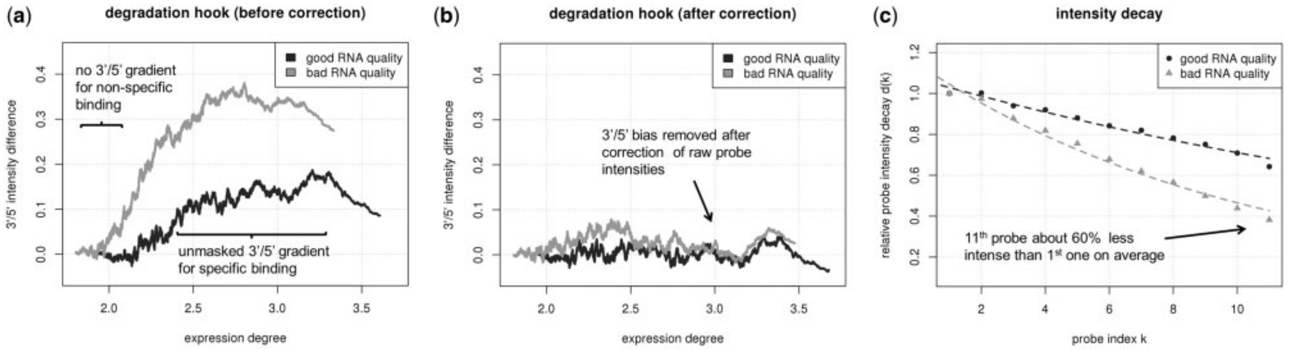

Fig. 1. Degradation hook plots referring to strongly and weakly degraded RNA taken from Archer *et al.* (2006) before [panel (**a**)] and after [panel (**b**)] correction using `AffyRNADegradation`. The height of the hook curve increases with increasing degradation level. Panel (**c**) shows the respective probe positional decays d(x) as plotted by the `AffyRNADegradation` package: the worse the RNA quality, the steeper is the respective decay

degradation hook-plot, shown in Figure 1a and b, displays this 3′/5′ intensity difference in dependence on the mean logged probe intensity approximating the expression degree of the respective gene. Cross-hybridization of partly matching targets of other genes causes nearly equal intensities for weakly expressed genes (Binder, 2006). With increasing expression competitive binding of specific targets progressively unmasks their actual 3′/5′ gradient, until probe saturation sets in. Desirable would be equal intensities for 3′ and 5′ probes for all expression levels. The maximum height of the hook-plot reflects the relevant 3′/5′-intensity gradient of the selected array enabling the unbiased comparison of differentially expressed genes under variable RNA quality.

The hook-plot is accessible using the `PlotDegradationHook` function in the package. A complementary representation is the Tongs Plot shown in the Supplementary Material and accessible using the `PlotTongs` function.

2.2 Estimation of the RNA quality of a sample

One should only use specifically hybridized probes for estimation of RNA quality because of the 3′/5′ gradient of the intensity as a function of the expression degree. For these probes, we compute the mean probe intensity separately for each probe index $k = 1\ldots 11$ starting from the 3′ end of the target transcript. Figure 1c shows the resulting probe positional intensity decay after normalization with respect to the mean intensity for the first probe $k = 1$. Alternatively, the intensity decay can be calculated as a function of the distance L of the probes given in units of nucleotides from the 3′-transcript end (not shown).

We determine the decay-length parameter d from the mean intensity decays of all specifically hybridized probes. It provides an accurate estimate for the RNA quality of a particular array hybridization improving other array-based metrices (Fasold and Binder, 2012). The $d(x = k,L)$ plot is available via the `PlotDx` function, and the RNA quality estimate is available via the `d` function in the `AffyRNADegradation` package.

2.3 Correcting the RNA quality bias

Differences in RNA quality and the resulting probe positional intensity decay are technical artifacts that can affect expression measures and the results of differential expression analysis. Microarray experiments are often subject to such RNA quality variation (Upton *et al.*, 2009).

We here aim at removing the systematic differences in probe positional intensities between different conditions. Figure 1a shows two such conditions in the example data relating to degraded transcripts due to increased presence of RNases not removed by RNeasy treatment. `AffyRNADegradation` first estimates specific probes based on the degradation hook-plot described above. It then uses a correction function that reverses the probe positional intensity decay d(x) after applying the expression level dependency of the hybridization mode (details are given in the Supplementary Material). Optionally, the correction can be performed based on probe indices k as well as probe distances L. Differences between both options are discussed in the Supplementary Material and in (Fasold and Binder, 2012). Figure 1b shows the degradation hook after correction using probe indices k: The 3′/5′ bias is almost completely removed. Corrected probe intensities are available via the `afbatch` function.

2.4 Package usability

The `AffyRNADegradation` package extends the Bioconductor package `affy` and integrates well in a typical microarray analysis workflow. All calculations are performed directly on the AffyBatch object and carried out separately for each particular microarray hybridization in a single-chip approach. Our approach corrects the 3′/5′-bias on the level of raw probe intensities, which can afterward be processed with any method. The runtime is about 2 min and 3 min per sample for index and distance based corrections, respectively. Because each chip is processed independently, arbitrarily large data sets can be processed.

ACKNOWLEDGEMENTS

This publication is supported by LIFE—Leipzig Research Center for Civilization Diseases, Universität Leipzig.

Funding: This project was funded by means of the European Social Fund and the Free State of Saxony.

Conflict of Interest: none declared.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Universitaet Leipzig, Institut fuer Informatik/URZ, Bibliothek on January 29, 2013

REFERENCES

Archer,K.J. *et al.* (2006) Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models. *Biostatistics*, **7**, 198–212.

Binder,H. (2006) Thermodynamics of competitive surface adsorption on DNA microarrays—theoretical aspects. *J. Phys. Cond. Mat.*, **18**, S491–S523.

Fasold,M. and Binder,H. (2012) Estimating RNA-quality using GeneChip microarrays. *BMC Genomics*, **13**, 186.

Fleige,S. and Pfaffl,M.W. (2006) RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.*, **27**, 126–139.

Gautier,L. *et al.* (2004) affy-Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Ma,C. *et al.* (2006) In vitro transcription amplification and labeling methods contribute to the variability of gene expression profiling with DNA microarrays. *J. Mol. Diagn.*, **8**, 183–192.

Schroeder,A. *et al.* (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, **7**, 3.

Upton,G.J. *et al.* (2009) On the causes of outliers in Affymetrix GeneChip data. *Brief. Funct. Genomic. Proteomic.*, **8**, 199–212.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Universitaet Leipzig, Institut fuer Informatik/URZ, Bibliothek on January 29, 2013

Supplementary text

AffyRNADegradation: Control and correction of RNA quality effects in GeneChip expression data

Mario Fasold[1,2]* and Hans Binder[1,2]

[1] Interdisciplinary Center for Bioinformatics; Universität Leipzig, D-4107 Leipzig, Haertelstr. 16-18
[2] Leipzig Research Center for Civilization Diseases; Universität Leipzig, Germany

* Corresponding author: E-mail: fasold@izbi.uni-leipzig.de, fax: ++49-341-9716679

Table of contents

1. Tongs plot and degradation hook

We present two graphical representations that allow assessing the degradation of RNA-transcripts in a chip-specific fashion. These so-called ‘degradation hook’ and ‘tongs plot’ estimate the 3’-enrichment of the probes and thus their degradation level in dependence on the expression degree. They depict the mean intensity difference between two selected subsets of perfect-match probes taken from the 3’- and 5’-ends of each probe set, respectively,

$$\Delta = \Delta\Sigma_{s3'/s5'} \equiv \left\langle \Sigma_p \right\rangle_{s3'} - \left\langle \Sigma_p \right\rangle_{s5'} \quad \text{(degradation hook)}$$

$$\Delta\Sigma_s \equiv \left\langle \Sigma_p \right\rangle_s - \left\langle \Sigma_p \right\rangle_{pset} \quad \text{(tongs plot)} \quad ,$$

$$\text{with} \quad \left\langle \Sigma_p \right\rangle_s \equiv \frac{1}{3}\sum_{k=i}^{i+2} \log I_k^{PM}$$

as a function of the average logged intensity $y = \Sigma \equiv \left\langle \Sigma_p \right\rangle_{pset}$ of all probes within the probe sets which estimates the expression degree to a rough approximation. The subscript s= s3’, s5’ assigns the respective subsets of three consecutive probes from either the 3’ or the 5’ end of the probe set. Three probes are chosen from each side of the probe sets to ensure robust averaging over their intensities *and* to ensure sufficiently large differences between the averaged intensities of both subsets, and thus a proper tradeoff between robustness and sensitivity. Note that the number of three probes refers to about 1/3 to 1/4 of the size of each probe set of typically 11 probes for most array types. On the other hand, our choice is not crucial: Selecting subsets of two or four probes from both ends of the probe sets only marginally alters the results (data not shown).

Figure S 1 shows the tongs plot and degradation hook for the same two samples that were used also in the main paper. In general, the degradation hook is more suited to compare the degradation between different samples. The tongs plot reveals additional information such as an asymmetrical behavior of the 3’ and 5’ subsets.

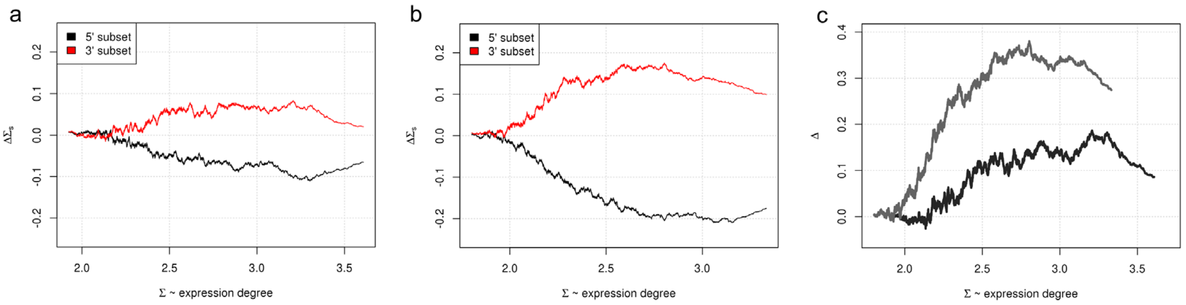

Figure S 1. Tongs plots (Panel a and b) and degradation hooks (panel c) of two array hybridizations using either weakly or strongly degraded RNA: With progressing degradation the ‘tongs opening’ (i.e. the maximum gap between the red 3’ and black 5’ branches) and the height of the hook increase. The two branches of the tongs plot and the two different hook curves converge at small abscissa values owing to the insensitivity of non-specific hybridization for degradation effects. The curves of both hybridizations are slightly shifted from each other in horizontal direction due to different scanner settings. The samples are taken from ref. (Archer et al., 2006) (samples VOV1_GOOD.CEL and VOV1_INHIBITED.CEL, respectively).

2. Probe positional intensity decays

Two main factors related to RNA quality potentially affect the intensities of the probes : (i) the distance of a probe relative to the 3’-end of the transcript, L (or, alternatively, the probe index in the probe set, k, which counts the probes in direction away from the 3’-end of the transcript) and (ii), the hybridization mode (Fasold and Binder, 2012). In the specific (S-) hybridization mode the probes bind amplified RNA (aRNA) fragments of complementary sequence originating from the mRNA transcripts which they intend to detect. In the N-hybridization mode the probes bind to aRNA fragments of partly complementary sequence originating however from other mRNA transcripts not referring to the interrogated gene. We normalize the intensity of probes at position x= k, L with respect to the intensity of probes located at the 3'-end to obtain the degradation index

$$d^{h}\left(x\right) = I^{h}(x) / I^{h}(3') \quad \text{with} \quad x = k, L \quad \text{and} \quad h = N, S \ ,$$

where $I^h(x)$ is the average perfect-match probe intensity of all probes with hybridization mode h=N or S and the same index k=1...11 (for x=k) or the same position L on the current array. For x=L, we use all probes in windows of +/-25 bases about the absolute positions L=25, 75, …, 575 (typically more than 95% of all probes are located within the range L=1..600). $I^h(3')$ is the average intensity level of the probes near the 3' end, i.e. $I^h(3')= I^h(k=1)$ and $I^h(3')= I^h(1<L<50)$ for x=k and x=L, respectively.

The degradation index due to non-specific hybridization is virtually constant $d^N(x) \approx 1$. The degradation index due to specific hybridization shown in Figure 1 of the main paper is described using an exponential decay of the form

$$d\left(x\right) = d^{S}(x) \approx \left(1 - d_{\infty}^{x}\right) \cdot \exp\left(-\frac{x - x_{0}}{\lambda_{x}}\right) + d_{\infty}^{x}$$

3. Correcting the probe positional intensity bias

The raw probe intensities of each sample are corrected as follows:

1. The degradation hook $\Delta\Sigma_{3'/5'}$-vs-Σ is calculated using all perfect-match probe intensities for the current array as described in section 1.
2. Probes are considered as specifically hybridized if the sigma-value of the respective probe set meets the condition $\Sigma_{3'/5'}^{max} - 0.4 < \Delta\Sigma_{3'/5'} < \Sigma_{3'/5'}^{max} + 0.2$ where $\Sigma_{3'/5'}^{max} = \arg\max\left\{\Delta\Sigma_{3'/5'}(y)\right\}$.
3. The decay function $d^S(x)$ (x=k, L) is calculated as described in section 2 using the subensemble of all specifically hybridized probes.
4. The mean fraction of probe intensities due to specific hybridization is estimated for each probe set as, $f^S(y) = \Delta\Sigma_{3'/5'}(y) / \Delta\Sigma_{3'/5'}^{max}$.
5. The correction function is calculated as weighted sum of the decay functions due to specific and non-specific hybridization where the latter one is simply set to unity, $d^N(x)= 1$, i.e.
$$C(x,y) = d^{S}(x) \cdot f^{S}(y) + d^{N}(x) \cdot \left(1 - f^{S}(y)\right) = d^{S}(x) \cdot f^{S}(y) + \left(1 - f^{S}(y)\right)$$
6. The biased probe intensities are then corrected using the inverse of the correction function, $I_{p}^{P,x-corr} = I_{p}^{P} / C(x,y)$.

Note that each probe intensity is rescaled according to value of the mean intensity decay at its position (x= k or L) and according to its hybridization mode as indicated by the abscissa-value of its probe set y. Consequently, probe intensities taken from the non-specific hybridization range remain uncorrected. With increasing degree of specific hybridization the probes are progressively scaled up with increasing distance from the 3'-end of the transcript. The maximum correction applies to probe sets in the S-hybridization range. MM probe intensities are scaled using the mean logged MM-intensity of the probe set as argument.

4. Positional information of the probes

The distances of the probes to the 3' end are obtained by aligning their sequences to the respective transcript sequence serving as target for the respective probe set as provided by Affymetrix. We have computed files containing probe positional information for a large number of GeneChip expression arrays. They are available via the website http://www.izbi.uni-leipzig.de/downloads_links/programs/rna_integrity.php.

These files are stored in R binary file format. The package documentation contains a section describing the contents of this file and explaining how the user can easily create and use custom probe location files, for example if he uses custom microarrays.

5. Choosing between absolute and relative probe positions

In the supplementary text of ref. (Fasold and Binder, 2012) we compare the two correction metrics based either on the absolute probe position ('L-correction') or on the relative probe position (index-based, 'k-correction') relative to the 3′ transcript end. The k-correction applies the same positional factor to all probe sets. In consequence, the probe set-specificity of the correction is solely determined by the degree of specific hybridization. Contrarily, the L-correction applies a specific factor to each probe-set depending on the particular location of its probes. Comparison of both correction methods shows that probe sets located on the average nearer to 3′-end of the transcript are corrected to a less degree using their absolute position than probe sets located more distant from the 3′-transcript end. Hence, the L-correction is more specific with respect to each particular probe set. On the other hand, the k-correction is more robust with respect to outliers.

We recommend use of absolute probe positions to cope with the effect of differently distributed probes. In practice the intensity changes due to index-based and position-based correction differ only slightly with, in general, small differences in the resulting expression values.

6. References

Archer,K.J. et al. (2006) Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models. *Biostatistics*, **7**, 198-212.

Fasold,M. and Binder,H. (2012) Estimating RNA-quality using GeneChip microarrays. *BMC genomics*, **13**, 186.