fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays

Matthew N. McCall¹*, Harris A. Jaffee² and Rafael A. Irizarry²

¹Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 265 Crittenden Blvd., CU 420630, Rochester, NY 14642 and ²Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA

ABSTRACT

Summary: Frozen robust multiarray analysis (fRMA) is a single-array preprocessing algorithm that retains the advantages of multiarray algorithms and removes certain batch effects by downweighting probes that have high between-batch residual variance. Here, we extend the fRMA algorithm to two new microarray platforms—Affymetrix Human Exon and Gene 1.0 ST—by modifying the fRMA probe-level model and extending the fma package to work with oligo ExonFeatureSet and GeneFeatureSet objects.

Availability and implementation: All packages are implemented in R. Source code and binaries are freely available through the Bioconductor project. Convenient links to all software and data packages can be found at http://mnmccall.com/software

Contact: mccalm@gmail.com

Received on June 26, 2012; revised on September 12, 2012; accepted on September 23, 2012

The majority of methods for the preprocessing and analysis of microarray gene expression data rely upon the simultaneous analysis of multiple arrays (Hochreiter et al., 2006; Irizarry et al., 2003; Li and Wong, 2001). However, most traditional preprocessing methods struggle with modern microarray applications, such as large meta-analyses, because data preprocessed separately cannot be combined without introducing artifacts (McCall et al., 2010; Ramasamy et al., 2008). Furthermore, clinical applications necessitate the analysis of individual arrays, and datasets that grow incrementally must be preprocessed each time a new array is added.

Frozen robust multiarray analysis (fRMA) (McCall et al., 2010) addressed these challenges by implementing a modified version of the RMA algorithm (Irizarry et al., 2003). Additionally, by modeling probe-specific variances, fRMA showed improved precision of gene expression estimates and reduced susceptibility to batch effects (McCall et al., 2010; McCall and Irizarry, 2011). The fRMA algorithm was initially implemented on two of the most widely used microarray platforms—Affymetrix GeneChip Human Genome U133A and U133 Plus 2.0. Since then, it has been implemented for several other platforms.

In 2007, Affymetrix released two new microarray platforms—Human Exon 1.0 ST (HuEx) and Human Gene 1.0 ST (HuGene). In contrast to previous platforms that targeted the 3’ end of transcripts, these new platforms contain probes for each exon. This design change allowed researchers to assess exon-level expression and detect alternative splicing. However, it also posed a challenge to those who wanted to use these arrays to assess gene expression using the same preprocessing algorithms that were designed for the previous generation of Affymetrix microarrays. Specifically, the majority of preprocessing algorithms assume that each probe within a probeset was designed to measure the expression of the same target transcript; however, when a probeset is composed of probes targeting different exons, this assumption may be violated due to alternating splicing. This is particularly problematic given that splice variants are estimated to occur in 35–59% of genes (Modrek et al., 2002).

By summarizing probes at the exon level, one revalidates the assumption that each probe within a probeset is measuring the same target. This is more feasible for the HuEx platform, which often has four probes per exon, than for the HuGene platform, which contains fewer probes (roughly 35% of exons are targeted by only one probe). However, the small number of probes per probeset limits the ability to generate robust estimates of expression.

To address these limitations and aid researchers seeking to assess gene-level expression using HuEx or HuGene arrays, we have implemented a modified version of the fRMA model for gene-level summarization:

\[ Y_{ijkl} = \theta_n + \psi_{ln} + \phi_{jn} + \gamma_{kn} + \epsilon_{ijkl} \]  

\[ \text{Var}(\gamma_{ijkl}) = \tau_{jm}^2; \text{Var}(\epsilon_{ijkl}) = \sigma_{jn}^2, \]

where \( Y_{ijkl} \) represents the log2 background corrected and normalized intensity of probe \( j \), targeting exon \( l \), of gene \( n \) on array \( i \) in batch \( k \). Identical to the standard fRMA model, \( \theta_n \) represents the expression of gene \( n \) on array \( i \) and is the parameter of interest. Here, \( \psi_{ln} \) represents the global probe effect for the \( l \)th probe targeting the \( k \)th exon of gene \( n \), and \( \phi_{jn} \) represents the exon effect for the \( h \)th exon of gene \( n \). These parameters are both constrained to sum to zero within exon and gene, respectively. Finally, \( \gamma_{ijkl} \) is a random effect representing the batch-specific change in the global probe effect. This model is fit as described in McCall et al. (2010) with an additional step to estimate the exon effects, \( \psi_{ln} \). For a new array, gene-level expression estimates are obtained as robust-weighted averages of the probe- and exon-effect adjusted expression values.

By using a large biologically diverse database of microarrays from a large number of different laboratories spanning several years, the fRMA algorithm is able to differentiate between...
outliers and probes that show a consistent susceptibility to batch effects. These batchy probes are downweighted during summarization to minimize their effect on expression estimates. HuEx and HuGene arrays add an additional layer of complexity—when summarizing at the gene level, a probe may show high between-batch residual variance due to either batch effects or alternative splicing (Fig. 1). The former should be downweighted, whereas the latter may contain highly interesting biological information that could be captured by subsequent analysis of residuals, such as those proposed in Robinson and Speed (2009). For this reason, even when summarizing to the gene level, a probe may show high residual variance even for exons targeted by multiple probes. For single-probe exons, it is impossible to assess residual variance at the exon level and, therefore, impossible to distinguish between batch effects and splice variants.

The two versions of the fRMA algorithm described earlier are implemented in the frma package and take advantage of the raw data structures implemented in the oligo package (Carvalho and Irizarry, 2010), allowing greater control over the level of summarization. Specifically, this is handled by the target argument passed to the frma function. The frozen parameter vectors for HuEx and HuGene arrays were created using 240 arrays from 48 batches and 1005 arrays from 201 batches, respectively. Here, a batch is defined as a unique tissue type/experiment combination. The frozen parameter vectors are stored in the huex.1.0.st.v2frmavecs and hugene.1.0.st.v1frmavecs annotation packages.

The frmaTools package (McCall and Irizarry, 2011), which allows users to create their own frozen parameter vectors, has also been updated to work with oligo GeneFeatureSet and ExonFeatureSet objects. This allows users to create custom vectors for the HuEx and HuGene platforms and to implement fRMA on other Affymetrix Exon and Gene ST platforms that are not currently supported.

ACKNOWLEDGEMENTS

The authors thank the maintainers of GEO and ArrayExpress for making the data publicly available, Marvin Newhouse and Jiong Yang for helping manage the data and the members of the La Calestienne Meeting, especially Hinrich Gohlmann and Willem Talloen, for their helpful discussions.

Funding: This work was funded by National Institutes of Health (CA009363 to M.N.M.), National Institutes of Health (GM083084, RR021967 and GM103552 to H.A.J.) and partially funded by National Institutes of Health (GM083084, RR021967 and UL1RR025005 to R.A.I.).

Conflict of Interest: none declared.

REFERENCES


