Microarray data quality assessment for developmental time series

Shiqin Helen Guan[1], Jie Zheng[4] and Julia Brettschneider[1,2,3]

[1] Mathematics / Statistics Dept., [2] Community Health / Epidemiology Dept.
[3] Cancer Research Institute, Div. Cancer Care / Epidemiology, [4] Biology Dept.,
Queen's University, Kingston, Canada

1 Introduction

While microarrays have become a routine-tools in modern molecular labs, the scientific community came to recognize that findings based solely on microarray measurements were not always reproducible, and that studies with inconclusive results were quite common. Inter-lab and inter-platform comparisons have demonstrated this convincingly, eg. Wang *et al.* (2005). Statisticians have recently started to systematically address the question how to quantify the overall quality of large number of massive parallel measurements like the ones produced by microarray experiments. Brettschneider *et al.* (2006) laid out concepts for microarray data quality assessment data in the spirit of Shewhart's statistical quality assessment philosophy. In particular, they introduce visual tools for assessing the spatial quality of short oligonucleotide microarrays, and numerical measures (NUSE, RLE) for the quality the pre-processed microarray data. The approach is derived based on the RMA model for microarray gene expression value estimation by Irizarry *et al.* (2003).

A straight forward application of these quality assessment methods on developmental time course data indicates particularly poor quality for the data from all the arrays at the beginning and at the end of each of the developmental series. We hypothesize that this is potentially due to a misinterpretation of biological variation as technologically caused noise. In fact, using NUSE and RLE for quality assessment is based on the assumption that the majority of genes is not differentially expressed. Developmental microarray data, however, differs in just this aspect from most other kinds of microarray data. The fraction of non-differentially expressed genes is much higher and it is inhomogenous over time. Another concern is wether the default RMA model we have generally been using in quality assessment is a good fit for the data. It does not have a factor for developmental stage, which may be too simplistic. On the other hand, microarray time course experiments usually provide no or only a small number of technical replicates, which makes fitting more complicated models a questionable endavor.

This paper aims at improving our interpretation of the microarray quality measures in the case of developmental experiments. We suggest and test alternative quality assessment strategies. geared toward developmental data. The first approach uses housekeeping genes selected using *a priori* biological knowledge. Restricting the quality assessment to this subset would allow to unambiguously attribute deviations in NUSE and RLE scores to technological causes. Our second approach uses a similar rationale, but uses statistical analysis to select the genes with low temporal variation. We further investigate whether model-misfit contribute to the poor quality we observed in the data from in the first and last few chips of each series. The small number of technical replicates is overcome by grouping by pooling data from several subsequent time points.

2 Material and Methods

Data. We use the data from `www.fruitfly.org/cgi-bin/ex/insitu.pl` created by Tomancak *et al.* (2002). Wild type fruit flies (Canton S) were split into 12 population cages and allowed to lay eggs after aged for 3 days. Embryo samples were collected every hour for consecutive 12 hours starting at 30 minutes. Embryonic stages were examined by morphological markers. The same procedure was conducted on 3 days yielding 3 biological time course replicates (A, B, C). RNA was extracted, labeled and hybridized to short oligonucleotide microarrays (Affymetrix) measuring 14010 genes with 11-20 probes per probe set. Series A and C series were hybridized on a day other than the one for series B. Each chip was named as the combination of the replicate label and the time point label 01, 02,...,12.

Quality measures. Microarray data quality is assessed by statistical quality assessment tools introduced in Brettschneider *et al.* (2006) using Bioconductor R-package `affyPLM` described in Bolstad (2005). The approach uses the RMA model by Irizarry *et al.* (2003). For a fixed probe set, the background adjusted and normalized intensities y_{ij} of probe j and microarray i are modeled as $\log_2 y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}$, with α_j a *probe affinity effect* (with zero-sum constraint), μ_i the log scale expression level for array i, and i.i.d. centered errors ε_{ij}. The use of an iteratively re-weighted least squares algorithms delivers a robust probe set expression estimate for this probe set for each array using the inverses w_{ij} of the residuals to down-weight malfunctioning probes. The *Normalized Unscaled Standard Error (NUSE)* of a fixed probe set is defined as the inverse of the total probe weight $\sum_j w_{ij}$ for array i normalized by that same expression across all arrays used in the experiment. The *Relative Log Expression (RLE)* of a probe sets in array i is computed as the log ratio of its expression value in that array and its median expression in a hypothetical reference array. The latter is usually obtained as the median of all arrays in that batch, constructed probe set by probe set. Under the biological assumption that the majority of gene is not differentially expressed, the collective behavior of the NUSE values measure the quality of the data from this array, relative to the quality of the data from the other arrays from the same experiment and jointly analysed. A good quality array has a median NUSE around 1 and a small IQR. The collective behavior of the RLE values can also be used as a quality indicator. A good quality array has a median RLE around 0 and a IQR.

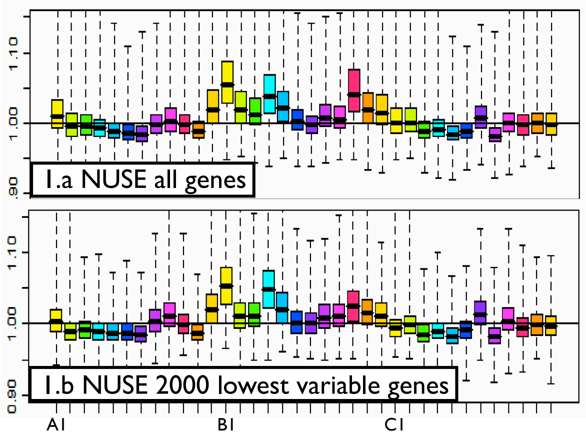
Alternative quality assessment approaches. The first three approaches use the RMA default model, but they in in terms of which genes are included in the quality assessment: (i) Genome-wide: all genes on the array; (ii) Constant genes picked by *a priori* biological knowledge: subsets of genes that are expected to vary little over time $-$ 54 housekeeping genes, and 87 temporarily stable genes according to Arbeitmann (2003); (iii) Constant genes determined empirically from our data: stratification of genes according to their degree of temporal variation using the empirical Bayes approach for microarray time course data developed by Tai and Speed (2006). Finally, approach (iv) uses a refined RMA model incorporating a factor for developmental stage. As the 3 replicates per stage may not suffice to fit such a model well, the 12 stages are divided in just 4 levels (1-3, 4-6, 7-9, 10-12).

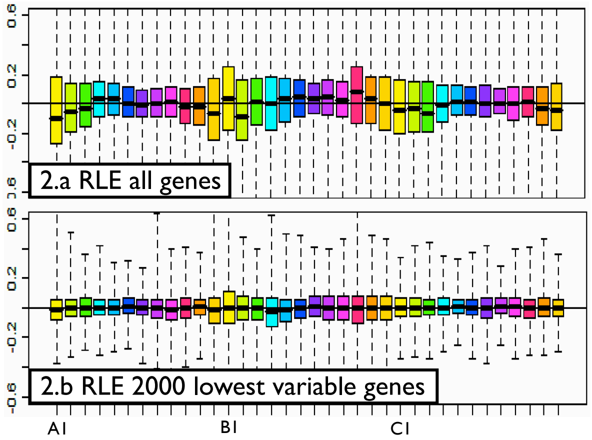
3 Results

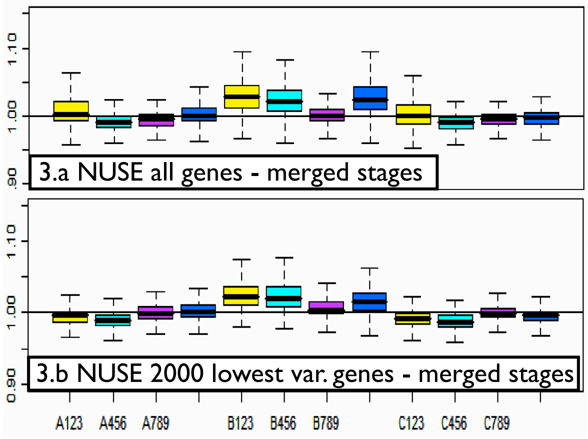
We begin with a straight forward application of the methods described in Brettschneider *et al.*, that is, whole genome based quality assessment with the default RMA model (Section 2: Approach (i)). The results are shown in Figures 1a and 2a with boxplots corresponding to each array. Arrays are ordered by series (A, B, C), and by stage within each series. Both RLE and

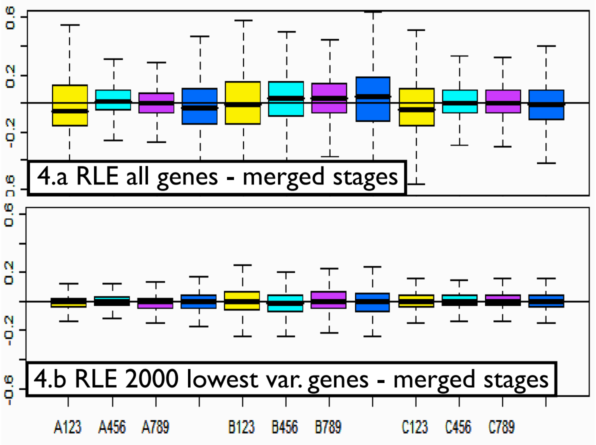
NUSE indicate that the quality of series B is generally lower than for series A and C. This most likely is due to the different hybridization date of series B. Note that the data has been quantile normalized before the quality assessment, which should take care of most overall differences between the series such as scanner calibration. Within all three series the middle stages seem to have produced better data quality. Most outstanding are the median NUSE and the IQR RLE of the first 2-3 arrays and the last 1-2 arrays of each series.

We wondered wether this is merely a reflection inconsistent degrees of biological variation between different developmental stages, rather than a sign of poor quality. Difficulties in stage assignment at the beginning of the series, and confusion between genes stimulated by the embryo itself and genes stimulated by the mother may contribute to further inconsistencies. We hypothesized that a restriction to genes expected to be constant would eliminate this problem. In Approach (ii) we restrict the quality assessment to genes known in the literature to be temporarily stable. It turned out, however, that the quality assessment restricted on two such sets (see 2) delivered results every similar to what we got in Approach (i).









For Approach (iii), we first determined the 2000 least changing genes based on a Hotelling statistics ranking, then we restrict the quality assessment to those 2000 genes. The NUSE distributions seem to be more or less unaffected by this selection (Figure 1b). The RLE distributions, however, react strongly to this restriction; the differences between the RLE distributions in the middle stages on the one hand and the early/late stages on the other hand are far less pronounced for this subset of genes (Figure 2b).

Consistent with the other approaches, (iv) also reveals that Series B is of lower quality than A and C (Figure 3a, 4a). The differences between the middle stages and the other stages are still there, but both less consistent and less pronounced. This may to some extend be due to the fact that the strongest outlier arrays have been pooled together with more normal arrays, but better model fit certainly plays a role. Restriction on the 2000 low varying genes makes nearly no difference for the medians of the NUSE distributions, as we saw in Approach (iii) as well. The IQRs of the NUSE distributions, however, are much smaller than they were in (iii), and they do shrink even a little more when restrict the assessment to low varying genes. This indicates an improved model fit. For the RLE distributions as well, the IQRs are smaller than in (iii), which confirms the better fit. Similar to the behavior of the RLE in (iii), there is a big difference in IQR RLE when we restrict the assessment to low varying genes.

4 Discussion

The interpretation of standard microarray data quality measures such as genome-wide NUSE and RLE appears to be problematic in the case of time course data. We suggested and tested alternative ways to assess microarray data quality for this data type. Involving *a priori* biological knowledge did not solve the problem. Stratification of genes by a statistical time course analysis, however, shed light on the issue. While the NUSE did not react to the stratification, the IQR RLE decreased a lot when restricting the quality assessment to a subset of the lowest variable 2000 genes. This inconsistency between NUSE and RLE is remarkable. Usually, despite their different construction, they give very similar results in quality assessment. Fitting a model with a stage factor indicates that some of the observed noise may be attributed to model misfit. Simulation studies and analysis for more developmental data sets are necessary to generalize and deepen our understanding of these issues.

Acknowledgement. The authors thank François Collin and Terry Speed for helpful comments about statistical and biological issues regarding this data set, and Pavel Tomancak for sharing details about experimental conditions.

References

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W. and White, K.P. (2002). Gene expression during the life cycle of *Drosophila meanogaster. Science,* 297:2270-2275.

Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A. and Speed, T.P. (2005). Quality Assessment of Affymetrix GeneChip Data in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R., Carey V., Huber W., Irizarry R., and Dudoit S. (Eds.), Springer.

Brettschneider, J., Collin, F., Bolstad, B. and Speed, T.P. (2006). Quality assessment for short oligonucleotide data. To appear in *Technometrics* (with discussion).

Irizarry, R., Bolstad B.M., Collin, F, Cope, L.M., Hobbs, B. and Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research,* 31(4): e15.

Tai, Y.C. and TP Speed, T.P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics,* 34(5).

Wang, H. and He, X. and Band, M. and Wilson, C. and Liu, L. (2005). A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics,* 6(1): e71.