Considerations in Making Microarray Cross-Platform Correlations

Manohar Kollegal*, Sudeshna Adak*, Richard Shippy**, Timothy Sendera**
**Computational Biology & Biostatistics Lab, GE Global Research, Bangalore, India*
***GE Healthcare Bio-Sciences, Chandler, Arizona, US*
{manohar.kollegal, sudeshna.adak}@geind.ge.com
{richard.shippy, timothy.sendera}@ge.com

Abstract

Comparison and integration of expression data derived from diverse microarray platforms is challenging. Factors affecting cross-platform comparison of microarray data include the choice of normalization method used, annotation differences, presence of splice variants, RNA degradation and probe distance from 3' end. A thorough evaluation of two commercial microarray platforms to determine an appropriate methodology for making cross-platform correlations is described here.

1. Introduction

Comparison and integration of data obtained from high-density oligonucleotide microarrays which are distinct in their manufacturing processes, choice of length of oligonucleotides, hybridization protocols, imaging and data analysis methods is challenging. Compounding to these cross-platform differences are the usual sources of variation existing within a given platform, which include effect of sample quality, optical noise and cross hybridization. The present study discusses factors influencing comparison of two commercial microarray platforms and provides guidelines for such comparison.

2. Background

Correlations between the different microarray platforms have been widely reported [1-5]. Shippy *et. al* [1] showed that by limiting the comparison data set to those genes which are uniquely represented and flagged as “Present” in the samples by the respective microarray platforms, the correlation improves. Similar results have also been reported while comparing long-oligo arrays and short oligonucleotide arrays [6].

3. Materials and Methods

Cross-platform data used in the present study [1] consists of microarray tests conducted using Affymetrix HG-U133 GeneChip® (A & B chip) and GE CodeLink™ UniSet Human 20K Bioarray. Two samples, Human Brain and Pancreas, were analyzed using 5 technical replicates from each sample. A common set of 10835 genes occurring on both the platforms was identified using a common build of UniGene cluster IDs. The metric chosen for comparing cross-platform data is Differential Gene Expression defined as a log2 transformed ratio of the gene expression in Brain to the gene expression in Pancreas. The statistical analysis was carried out in R [7] and Bioconductor [8] package.

4. Correlation and Fold Change Ratio

Considering only the genes in the high fold change region (Fig. 1a) of the overall correlation plot, the cross-platform correlation is high (0.899 among 2428 genes). Making use of the platform specific quality calls and considering only the concordantly ‘Present/Good’ genes, we find that the correlation increases to 0.953 in a reduced subset of 420 genes (Fig.1b). In the low fold change ratio regime by using the set of concordantly present genes, the correlation increases by 60% to 0.523 from 0.197.

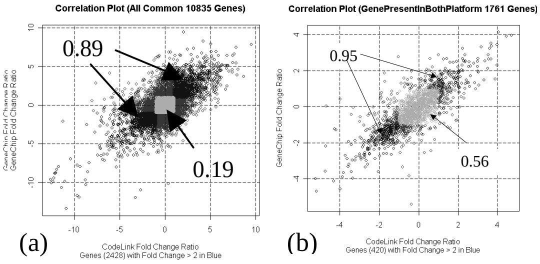

Figure 1(a) Correlation considering all genes (b) Correlation considering genes above noise

In the Discrepant Set region, amongst the key factors, which possibly influence cross-platform correlation, is the presence of splice variants. To understand this better, the genes in this set, need to be carefully studied and is currently being investigated.

5. Correlation and Sample Quality

Communication with the experimental team [1] revealed that the total RNA in Pancreas sample was partially degraded relative to the Brain sample, when observed on the Agilent 2100 Bioanalyzer. It was observed that by choosing a gene subset that was flagged as concordantly "Present/Good" in Pancreas, a higher correlation (0.81) in a larger set of genes (2153) was obtained than the 'concordantly present' case (0.78 among 1761 genes). This indicates that by choosing the 'Present/Good in Pancreas' subset of genes, we are able to remove system noise better than the 'concordantly Present' category.

6. Correlation and Normalization Methods

We next examine if the normalization methods that improve the within-platform replicate correlation in one microarray platform also help in improving the cross-platform correlation. The GeneChip® predicted fold change measures computed using different normalization methods (MAS5, dChip, RMA, GCRMA) are compared against the median normalized CodeLink™ fold change measures. Figure 2 shows the correlation existing between CodeLink™ and GeneChip® when we consider the set of all genes (10835), concordantly 'present' genes (1761) and genes (2153) marked 'present' in Pancreas sample only. The graph indicates that the cross-platform correlation does not improve by using the model-based normalization methods for GeneChip®.

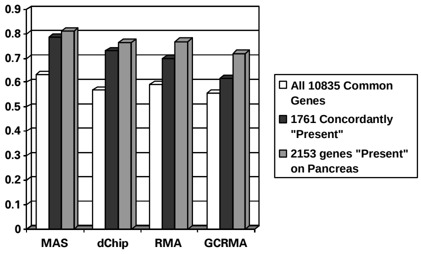

Figure 2 Correlation and Normalization Method

7. Correlation and Probe Distance

In general, cross-platform correlation increases as we select subset of CodeLink genes, which are closer to the 3' end. This indicates that in the case of moderately degraded samples, cross-platform correlation is also influenced by the choice of common probes and their distance from the 3' end.

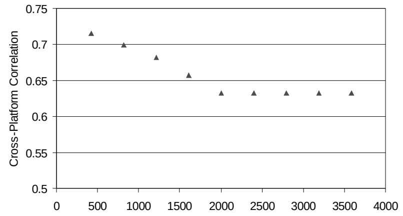

Figure 3 Correlation increases with subset of probes closer to 3' end

8. Conclusions

- Cross-platform correlation is high (0.9) among genes with high fold change (FC > 2) values and low at low fold change (FC < 2) values.
- Platform specific quality calls improve cross-platform correlation.
- Genes with signals, which are above the system noise level, yield higher cross-platform correlation.
- Sample RNA quality is quite likely to affect different genes differently. It is therefore desirable to identify genes/probe-pairs that are affected by RNA degradation so that a different procedure can be used to handle such genes during normalization or cross-platform comparison.
- Most platforms have their gene probes at various distances from the 3' end. It is quite likely that improved cross-platform correlation can be obtained by taking appropriate subsets of genes.

10. References

[1] Shippy R, *et. al.*, BMC Genomics, 5:61, 2004

[2] Jarvinen *et. al*, Genomics 83, pp 1164-1168, 2004

[3] Mecham, *et. al.*, Nucl. Acids Res., Vol. 32, No. 9, e74, 2004.

[4] Tan *et. al.*, NAR. Oct 1;31(19):5676-84, 2003

[5] Yauk *et. al.*, NAR. Aug 27;32(15):e124, 2004

[6] A. Barczak *et. al.*, Genome Research, 13:1775-1785, 2003.

[7] R: A language and environment for statistical computing, http://www.R-project.org

[8] Robert C Gentleman, at. al., Genome Biology, Vol 5, 2004, R80.