

A probe-to-transcripts mapping method for cross-platform comparisons of microarray data taking into account the effects of alternative splicing

Chunlei Wu¹, Jeffrey S. Morris¹, Keith Baggerly¹, Kevin R. Coombes¹, John D. Minna², Li Zhang¹

Keywords: microarray, cross-platform, comparison, transcript, alternative splicing

1 Motivation

Due to differences in hardware design and sample processing protocols, the gene expression quantifications obtained from different microarray platforms are often drastically different. Even for different versions of Affymetrix GeneChip microarrays, in which the differences are merely in the probe sequence selection, discrepancies are frequently observed from data that supposedly represent the same genes on the same samples. The lack of comparability poses a serious challenge in research projects that require analysis of microarray data from heterogeneous sources. Besides differences in sensitivity and specificity of the measurements on different platforms, we believe that the major cause is the failure to account for alternative splicing. By design, the effects of alternative splicing are ignored in most microarray platforms where each probe is assumed to match only a single gene transcript. It is obviously unreasonable in mammalian species, for which alternative splicing is prevalently observed [1], which may explain the discrepancies mentioned above.

2 Results

We have developed a new method to facilitate comparison of microarray data using different probeset definition. The method takes into account the effects of alternative splicing. Our approach is based on the recognition that in order to ensure concordant behavior of two different probes, it is essential that the two probe sequences match the same set of full length mRNA transcripts, which include all alternatively spliced variants. This realization leads to a new definition of how to map observed probe signals from one platform to another.

We first constructed a comprehensive library of full-length mRNA transcript sequences in the human genome by combining records in RefSeq (build 111504, human section)[2] and H-InvDB (version 1.7) [3] databases. For each probe sequence on HG-U133A and HG-U95Av2 arrays, all matched full-length transcripts were identified using the BLAST. The IDs of the transcripts with exact matches were collected to compose a matched-target list. By grouping the probes with the same matched target lists, we formed 23972 and 14148 probesets on the HG-U133A and HG-U95Av2, respectively. We call these probesets *Full-Length Transcript Based Probesets* (FLTBP). By matching the matched target lists of FLTBP on the two arrays, we found 9642 pairs of FLTBP that can be mapped between the HG-U133A and HG-U95Av2 for cross-platform comparisons.

¹Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 447, Houston, TX 77030, USA. E-mail: cwu@mdanderson.org

²Hamon Centre for Therapeutic Oncology Research, University of Texas Southwestern Medical Centre, Dallas, TX 75390, USA

We tested this method in a comparison of 28 paired samples using two versions of GeneChips, HG-U95Av2 and HG-U133A. Our method demonstrated marked improvement over the default mapping method from Affymetrix that ignores the effects of alternative splicing. (Figure 1,A-D) We also demonstrated that the best cross-platform consistency is observed when PDNN model [4] is used to quantify gene expression levels, when compared with RMA[5], dChip[6] and MAS5[7], which are commonly used alternative methods (Figure 1,E).

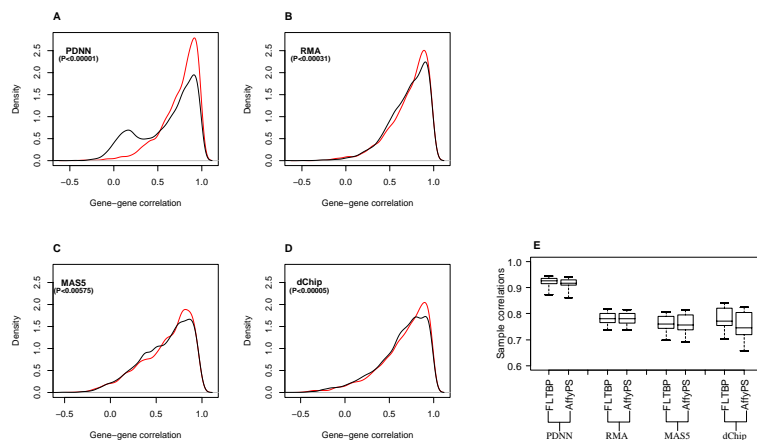


Figure 1: FLTBP/PDNN combination resulted in better cross platform consistency across different versions of Affymetrix GeneChip arrays. **(A-D)** Distribution of the gene-by-gene correlations between the paired measurements of two probesets across samples. The red line corresponds to the FLTBP mapping and the black line corresponds to the Affymetrix mapping. Each panel contains the correlation results using one of four quantification methods: A) PDNN, B) RMA, C) MAS5 and D) dChip. P-values shown in the plot indicate the significance of Kolmogorov-Smirnov test between two distributions. Better correlations are observed with FLTBP mapping than that with Affymetrix mapping, especially when using PDNN method (panel A). **(E)** Boxplot of sample-by-sample correlations between paired measurements on two different arrays across genes. For each method, sample-sample correlations are calculated using FLTBP mapping and Affymetrix mapping (AffyPS), respectively. Better correlations are observed when PDNN method is used.

References

- [1] Modrek B., Lee C. 2002. A genomic view of alternative splicing. *Nat Genet*, 30:13-19.
- [2] Pruitt K.D., Katz K.S., Sicotte H., Maglott D.R. 2000, Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 16:44-47.
- [3] Imanishi T., Itoh T., et al. 2004. Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biology* 2:856-875.
- [4] Zhang L., Miles M.F., Aldape K.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol* 21:818-21.
- [5] Irizarry R., Bolstad B., Collin F., Cope L., Hobbs B., Speed T. 2003, Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31:e15.
- [6] Li,C. and Wong,W. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* 98:31-36.
- [7] Affymetrix (2001) Microarray Suite User Guide, Version 5. Affymetrix, <http://www.affymetrix.com/support/technical/manuals.affx>.