Exon Array Background Correction

We expect to obtain similar performance in detection of differential changes for most applications on exon arrays using the background probe collection (BGP) as opposed to specific mismatch probes (MM). (For example, the GeneChip® Human Exon 1.0 ST Array has a “genomic” and an “antigenomic” background probe collection. See Appendix for more information on the background probe collections.)

In this whitepaper BGP probes are used to perform a PM-GCBG correction. PM-GCBG refers to the use of the median BGP intensity for BGP probes with the same GC content as the perfect match (PM) probe.

MM probes provide controls similar in sequence and location which are useful in removing bias; this takes one probe per informative probe. Because the MM probes hybridize to the perfect match (PM) target, when MM intensities are subtracted from PM intensities the PM response to true target slightly reduced. BGP probes provide a rough estimate of background based on coarsely modeled sequence, without taking into account any spatial variation. Use of BGP over MM probes requires roughly 50% less space on the array; thus use of BGP probes allows roughly a 2 fold increase in PM content.

Background estimates from BGP and MM probes may not be accurate for specific PM probes leading to biases in the background estimates results. We therefore look at the typical distribution of error in modeling background, as well as the downstream results on the ability to detect differential change. These results are based on a research array design using a pre-optimized version of the whole transcript assay.

We examine the performance of these two methods with a controlled experiment on a research chip using spiked-in transcripts. This experiment consists of 9 full-length clones which are spiked in at several different levels of relative abundance, with [3] replicates per level of abundance. The levels are arranged in a standard latin square design so that each experimental pool has some transcripts at each level of abundance, and every transcript occurs at all levels of abundance. In this experiment, the standard HeLa background is used as a complex background, and is not varied. This experiment was performed on an initial testing array design(“WTA-Sensor”) with a pre-optimized version of the whole transcript assay. protocol. This array design contained probes designed to complement every other base in the full-length transcripts. For analysis purposes, the transcripts were divided into twenty sub-regions each (simulating exons), and four probes were chosen to represent each sub-region, leading to twenty probe sets per transcript, or 180 probe sets analyzed in the data set used here.

In Figure 1, the distribution of multiplicative error for three different background models is displayed for a collection of background probes. By design, these should all be without specific hybridization, and hence an ideal background

model would exactly predict the intensity of these probes. What is displayed in figure 1 is a boxplot of the residuals log(I)-log(B) for each model. The first model is the log-scale error involved in approximating the background for probes of differing GC content by a single, universal background quantity, in this case, the median intensity. The second model sets the background estimate for a given probe to be the median of the corresponding bin of probes with similar GC-content, i.e log(I)-log(B(GC)). The last model sets the background estimate to be the corresponding MM probe. Both MM and GC-bin medians closely follow the behavior of probes across a wide range of GC-content, although GC-bin medians are slightly biased. Using a single global estimate of background has an unacceptably large distribution of errors in background prediction, however, GC-bin medians are sufficiently low error to be considered as replacements for specific MM probes.

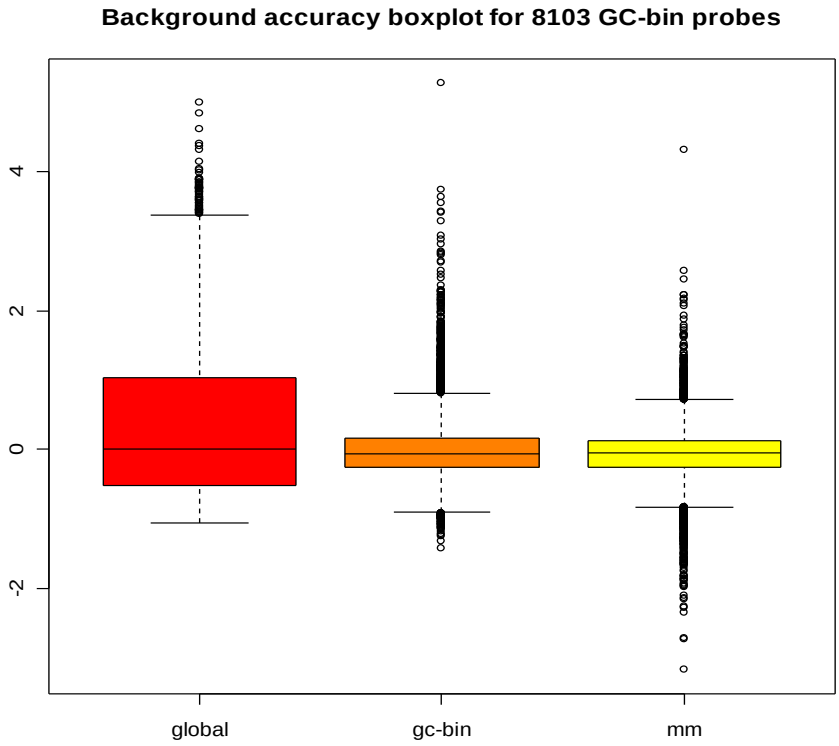

Figure 1: The log-scale error in approximating background by three techniques: global, gc-bin averages using BGP, and individual mismatches (MM) is shown here.

Figure 2 shows the mean-squared log-scale error resulting from each of the three considered models. Use of a single, global background estimate leads to large mean-square-error, and so is not acceptable for background estimation. Despite the very similar inter-quartile ranges for both the gc-bin background model residuals and the mm model residuals, the mean-square error is larger for gc-bin

probes. However, the known tendency of mismatched probes to hybridize to the true target suggests that gc-bin background estimation may be sufficient for background estimation in the context of detecting differential changes in signal.

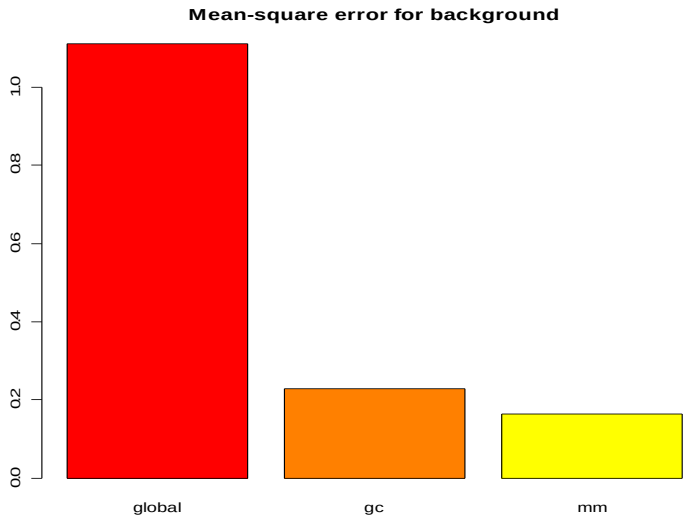

Figure 2: Despite the similarity in the boxplots, mismatched probes (MM) are still a better approximation in mean-squared error (log-scale). However, since there is some loss of signal due to MM hybridization to the desired target, this extra variation does not greatly impact signal accuracy.

Using the same data set, performance was evaluated for the various background estimation methods. In this case, performance is measured by the ability to detect differential change between paired experiments. In the figure below we show receiver operator curves (ROCs) based on thresholding a t-like statistic between the different spike concentrations. Use of BGP over MM probes does not adversely affect the ability to detect differential change. It was observed however that some probesets exhibited more positive bias with BGP (data not shown), indicating that coarsely modeled background with BGP does not completely remove bias from the data to the extent of the PM-MM method. This may cause problems in direct comparison of extremely different samples, which is not reflected in this experiment.

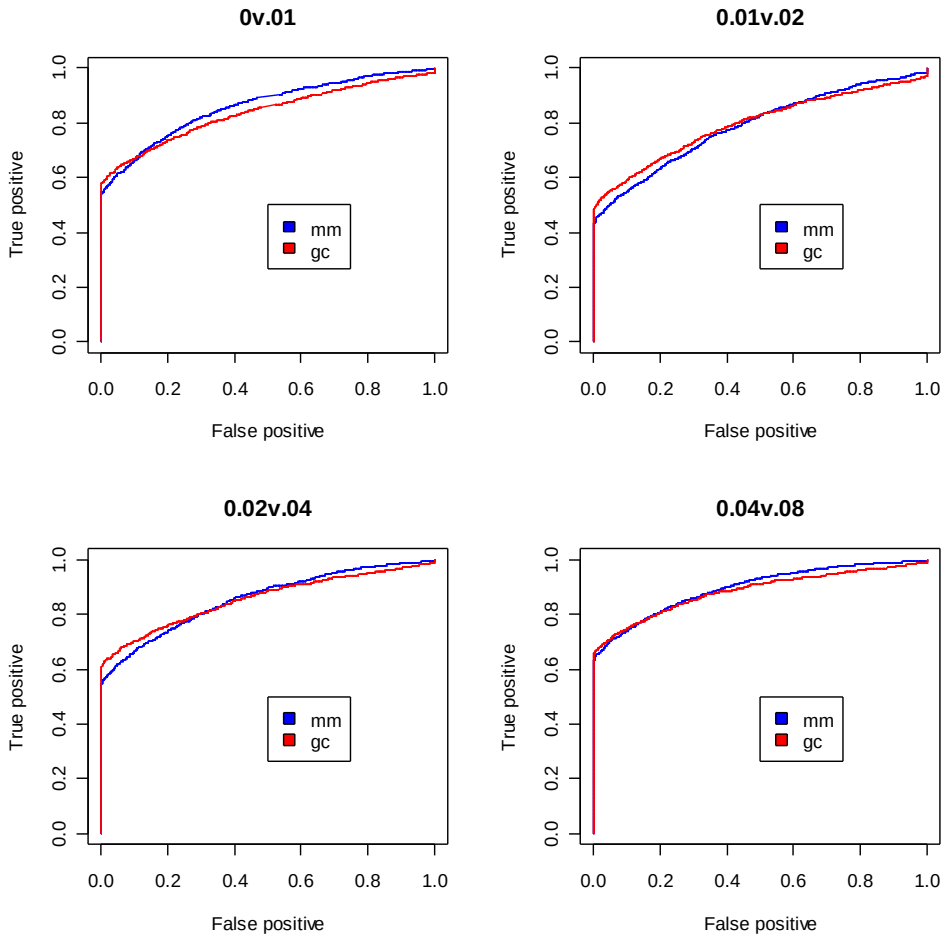

Figure 3: t-statistic based ROC curves for comparison between different conditions, here denoted as 0, .01, .02, .04 and .08. These are relative abundances before amplification, corresponding to 0, 1:300K, 1:150K, 1:75K, 1:37K. The ROC curves are averaged pairwise performance over all replicates, and are generated using only 4 probes. Note that these ROC curves are based on a testing array with an un-optimized whole transcript assay, and as such the absolute performance shown here (ie 60% sensitivity at near 100% specificity for 0 v 1:300K) is not indicative of the exon array system performance. See the GeneChip® Human Exon 1.0 ST Array Performance technote for system performance information.)

It is important to note that these results are based on an assay that generates DNA target, not RNA target. Thus the observations for the whole transcript assay with regard to BGP may not hold for the 3' IVT assay.

Appendix: Genomic and Antigenomic Background Probes

I. Genomic Background Probes

Background probes were selected from a research prototype human exon array design based on NCBI build 31. Mismatch probes were designed against Genscan Suboptimal exon predictions that were not supported by any other annotation source. When possible, 1000 mismatch probes were selected for each bin of GC content from 0 to 25..

II. Antigenomic Background Probes

The Antigenomic background probe sequences are derived from 15-mer root sequences that are not found in the human (NCBI build 34), mouse (NCBI build 32), or rat (HGSC build 3.1) genomes. These root sequences were extended by 5 bases on each end, where the added bases were drawn randomly from a GC distribution identical to the base 15-mer. There are probes for GC counts ranging from 2 to 24 per probe sequence. When possible, 1000 mismatch probes were selected for each bin of GC content.