

Two color array normalization based on intensity of spiked in controls

Svetlana Shchegrova¹, Paul Wolber¹, Petula D'Andrade¹

Keywords: oligonucleotide arrays, normalization, spike-ins, dye bias, dilution study

1 Introduction.

Microarrays with one and two color sample labeling are widely used to measure relative abundance of thousands of mRNAs at a time. Two-color approach has a number of advantages. The most important of which is that the experimental variations cancel out to a large degree in the final expression measure – the ratio of the two channel signals. However, before this ratio can be produced the intensity data from each channel must undergo some processing. In particular, a step called normalization has to be performed in order to remove the dye bias. Popular normalization methods are predominantly global. They include normalization to the mean/median and normalization using an invariant set of genes. The latter is implemented in Agilent Feature Extraction software under the name of rank-consistency method. dChip software described in [1] uses the same principle for one-color array processing. At the root of these approaches lies the assumption that the expression of the majority of genes on an array is unchanged. This might be justified in some cases however one can easily imagine a scenario where this assumption breaks.

We present an attempt to normalize the data using intensities of RNAs introduced into the sample at known concentrations, so called “spike-ins”. Spike-in normalization technique can be a powerful alternative to global normalization because i) it has more chances to work properly when data is skewed; ii) it improves consistency of data across arrays; iii) it makes possible comparison of two – color data with one-color data; iv) it allows absolute abundance evaluation. In this study a direct comparison is made between spike-in normalization and the rank-consistency method. A similar normalization approach was used in [2].

2 Experiment Description.

The experiments were done using Agilent Human 1A (v2) arrays that have over 18,000 human genes. We mixed HeLa and human Spleen polyadenylated RNA in 9 different proportions. The mixtures were labeled with Cy5 dye. Human Spleen was labeled with Cy3 dye and used as a common reference. The series included three replicates for each mixture. Synthetic transcripts with randomly generated 60-mer inserts were spiked-in in different concentrations into samples prior to labeling. The concentrations were chosen to produce a wide range of red to green signal ratios. A description of these synthetic targets can be found in [3].

3 Calculations and Results

Following basic assumptions we expect a linear response in spike-in intensities as a function of concentration. However, due to saturation effects a slight curvature is observed. The same observation was made in [2]. Therefore, we find it more appropriate to fit a line to the log-transformed spike-in intensities according to the formula $\ln(I) = a \ln(c) + b + E_j$, where I is

¹ Agilent Technologies, Inc., 3500 Deer Creek Road, Palo Alto, California, CA 94304, E-mail: Svetlana_shchegrova@agilent.com

