Supplementary Materials: Estimating the Proportion of Microarray Probes Expressed in an RNA sample

Wei Shi[1], Carolyn A de Graaf[1,2], Sarah A Kinkel[1,2],
Ariel Achtman[1],Tracey Baldwin[1], Louis Schofield[1,2],
Hamish S Scott[3,4], Douglas J Hilton[1,2] and Gordon K Smyth[1,5]

[1] *The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia*

[2] *The Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia*

[3] *Institute of Medical and Veterinary Science and The Hanson Institute, Adelaide, South Australia, Australia*

[4] *Adelaide Cancer Research Institute, The School of Medicine, University of Adelaide, South Australia, Australia*

[5] *The Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria, Australia*

1 Estimating expression proportions using different negative control probe intensities

Figure S1 shows the proportions of non-expressed probes estimated for each of the fifteen arrays in a dataset using MouseWG-6 V2 BeadChips (**Table 1**, dataset 3). Each line represents an array. For each array, the proportion of non-expressed probes was cal-

culated by testing on each background intensity (intensities from each negative control probe). The MouseWG-6 V2 platform encompasses 936 negative control probes on the array. Therefore, 936 estimates were calculated for each array. They were plotted in the order of increasing background intensities.

The estimated proportions are not stable when testing on negative controls which had very small intensities (the leftmost side of lines). The reason for this is that there are very few regular probes and negative control probes falling into this intensity range which makes it impossible to estimate the intensity distribution properly. Proportions are quite stable at the median intensity range. The median negative control intensity is used to estimate the expression proportion.

2 Proportion of detected probes using Illumina's detection p-values

Figure S2 shows the proportions of detected probes in each array for different platforms obtained by using Illumina's detection calls (detection p-values). The cut-off for the detection p-values is set to 0.01. Every platform has a smaller proportion of expressed probes compared to the same platform in Figure 2 in the main manuscript. The increasing pattern of expression proportions along the BeadChips versions in Figure 2 is also lost.

3 Estimating the expression proportion at gene level and transcript level

Multiple probes could be designed for the same gene/transcript in a microarray. To estimate the expression proportion at the gene level, we need to select one probe for each gene. The probe selected for a gene in this study is the one which had the largest average

expression value across all arrays in an experiment among the probes designed for the same gene. Results of estimating expression proportions at the gene level are shown in **Figure S3**.

Similarly, estimating the expression proportion can be performed at the transcript level (RefSeq genes). Results can be seen in **Figure S4**.

Probe annotation provided by Illumina is used in this analysis.

4 Numbers of genes expressed commonly and uniquely

The estimated expression proportions for the mixed samples and pure samples can be used to infer the numbers of genes expressed commonly and uniquely in the two pure samples. The proportion of genes uniquely expressed in one pure sample is the difference between the expression proportion estimated for the mixed sample and the expression proportion estimated for the other pure sample. For the in-house mixture data, the expression proportions for the mixed sample was calculated as the average of expression proportions for samples 88%:12%, 76%:24% and 50%:50%, which was 0.618. The expression proportion for the pure sample "MCF7" was calculated as the average of expression proportions for samples 100%:0% and 94%:6%, which was 0.594. The expression proportion for the pure sample "Jurkat" was 0.586 (average of its two replicate arrays). Therefore, the proportion of probes uniquely expressed in "MCF7" was 0.618-0.586=0.032 and the proportion of probes uniquely expressed in "Jurkat" was 0.618-0.594=0.024, which corresponded to 631 and 474 probes respectively. The proprotion of probes expressed in both samples can thus be calculated as 0.618-0.032-0.024=0.562, which corresponded to 11,088 probes.

These numbers can be computed similarly for the MAQC data. The proportions of probes expressed uniquely in pure samples "UHRR" and "HBRR" were 0.036 and 0.034 respectively. The proportion of probes expressed commonly is 0.704.

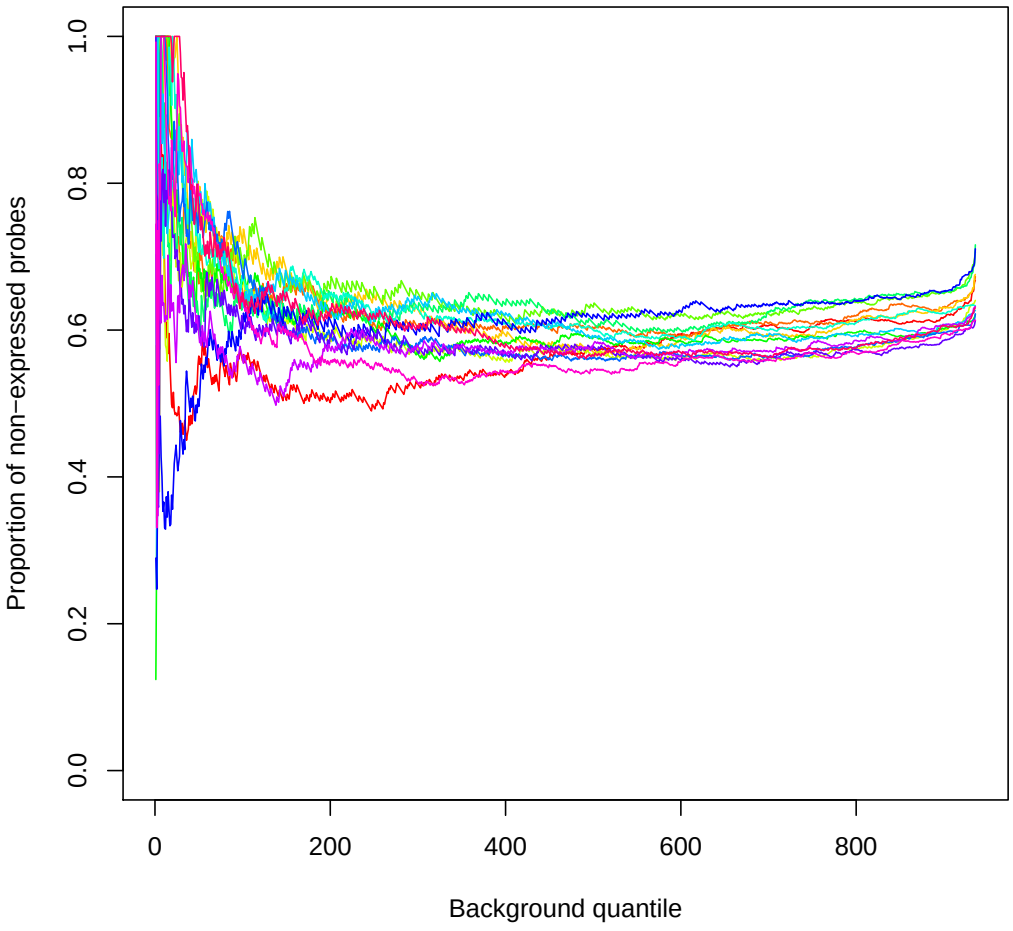

Figure S1: Proportions of non-expressed probes estimated from all negative control probes. Each line represents an array. Fifteen arrays were used. The array type is MouseWG-6 version 2.

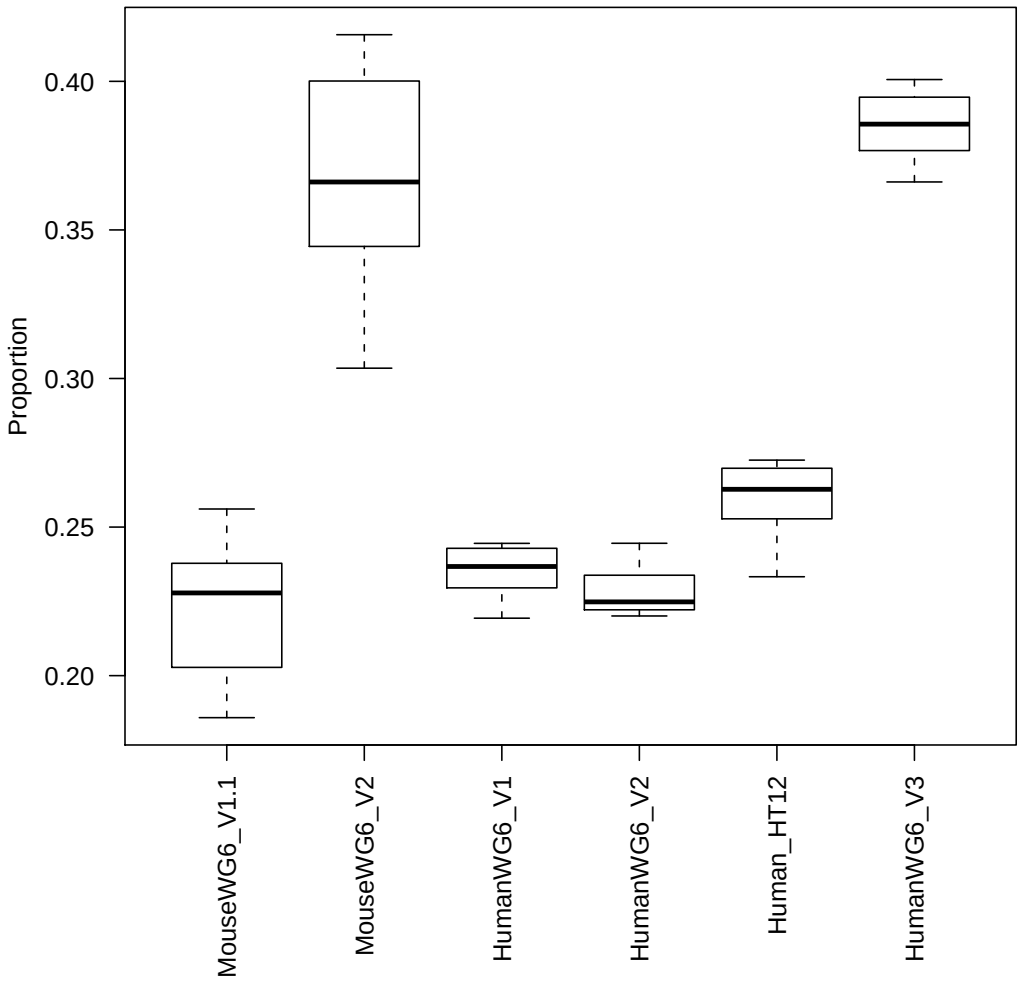

Figure S2: Use Illumina's detection p-values to get the proportion of detected probes for each BeadChip type. A p-value cut-off of 0.01 is used.

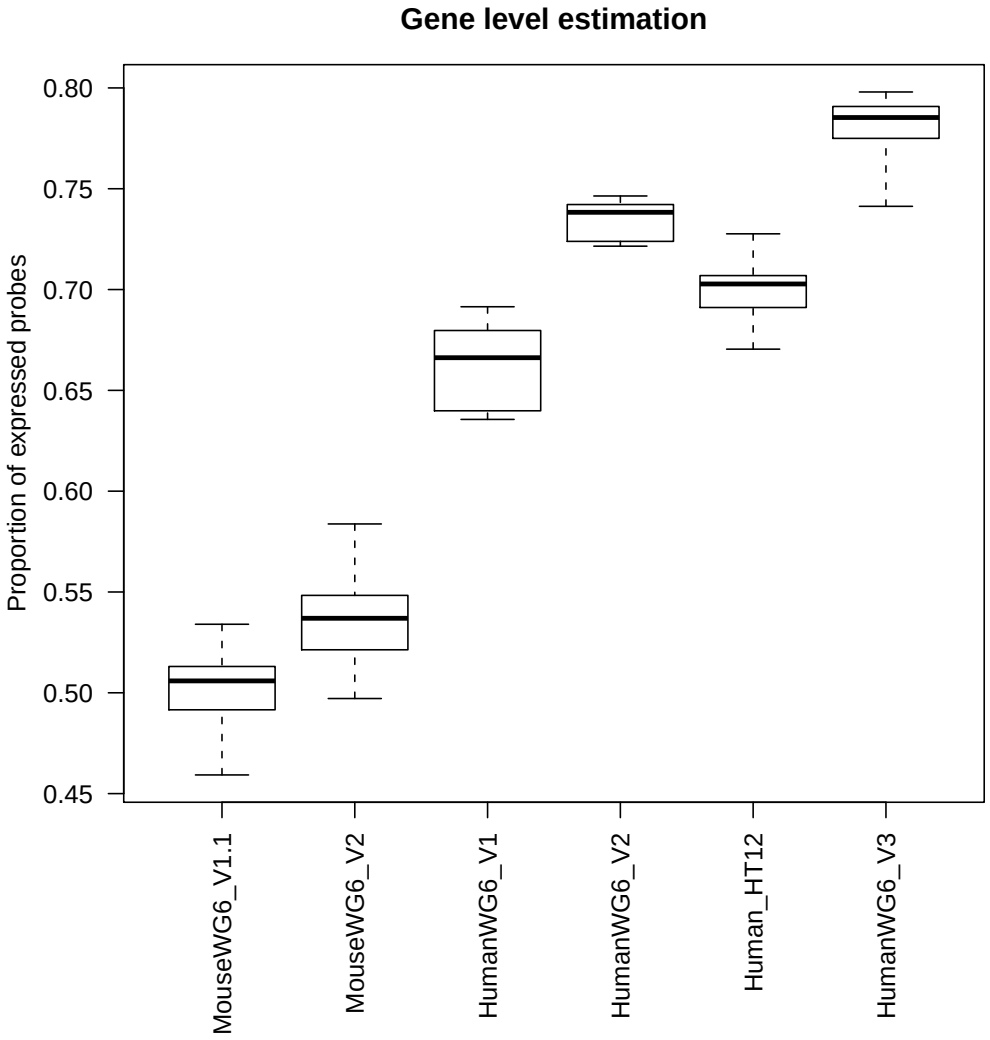

Figure S3: Estimating the expression proportion at gene level.

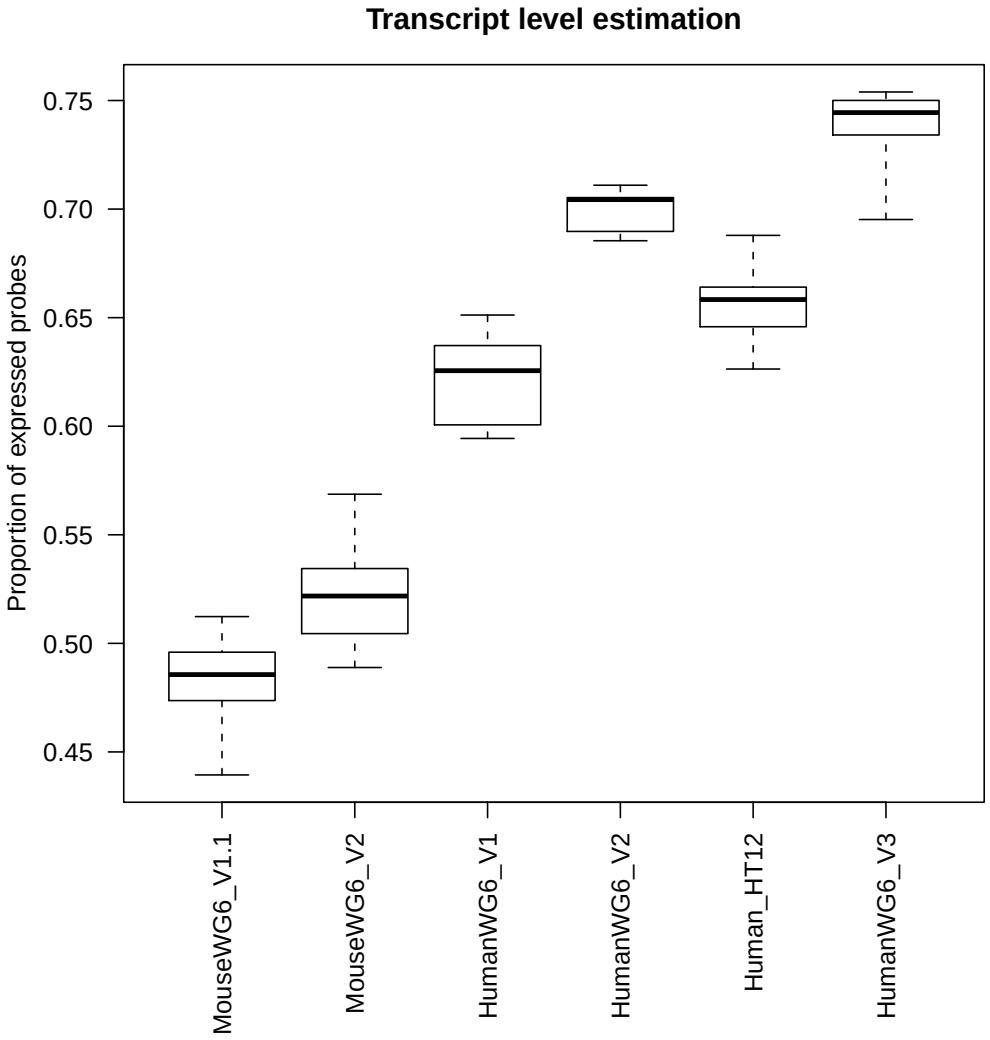

Figure S4: Estimating the expression proportion at transcript level.