Correcting for probe-design in the analysis of gene-expression microarrays.

Andy Lynch[1], Christina Curtis[2] and Simon Tavaré [1,2]

[1] Department of Oncology, University of Cambridge
[2] Department of Biological Sciences, University of Southern California

1 Introduction

The oligonucleotide probes on microarrays for gene expression are carefully designed so that their thermodynamic properties are consistent, in order that any bias due to differential efficiencies of hybridization might be minimized. However there are numerous other constraints such as the need to match (and be specific to) a gene, as well as possibly having a fixed probe length. Thus a perfect thermodynamic balance cannot be achieved. Also, since the labelling dyes can hinder hybridization, any thermodynamic balance achieved would not be robust to a change of the choice of dye, or method of adhering the dye to the RNA sample.

Aside from Affymetrix platforms (Zhang *et al.* 2003, Abdueva *et al.* 2006), such effects are not usually accounted for when analysing the results of microarray experiments. In particular with two-channel platforms, one might anticipate that any effects would cancel out across the two channels. Here we focus on the Agilent Human 1A Oligo Microarray (V2), a popular two-channel array. We comment on the design of its probes, and illustrate the ways in which those designs can bias downstream analyses before discussing possible remedies.

2 Our Data and the Human 1A Oligo Microarray (V2)

This Agilent microarray consists of $22,575$ locations arranged in a 105 x 215 grid. 422 spots are not reported, and $1,080$ contain control probes, leaving $21,073$ unique probes (all 60mers) that comprise the 'business part' of the array. Of these, the annotation we use suggests that $17,579$ map to areas of the genome that are transcribed, and of these $16,823$ to the autosomal chromosomes. It is on these that we focus, making particular use of the $2,686$ probes that have a non-unique target gene. In total, $1,305$ genes are targeted by more than one probe.

Whereas several manufacturers seek to have a tight distribution of GC content, as part of the thermodynamic control, Agilent have a very distinctive pattern for GC content (Figure 1, top left), with modes at 21, 27, 33 and 36 bases out of the 60. The distribution of GC content varies considerably between those probes ending in a G or C and those ending in an A or T, far beyond the natural constraints that this imposes (Figure 1, top centre). While there is a trend along positions 1 to 59 of the probes for increased discordance of GC content distributions depending on the GC status of that base, it is clear that the major effect is at position 60 (Figure 1, top right). The distribution of As, Cs, Gs and Ts is very much constant along the probes, save for the beginning and end positions (Figure 1, bottom).

The end base is known to be important in several models of the thermodynamics, and it is no surprise to see a) the change in base frequency at that point and b) the associated change in distribution of GC content. Less explicable are the change in base frequency at the first position and the increasing association with the GC content from positions 1 to 59.

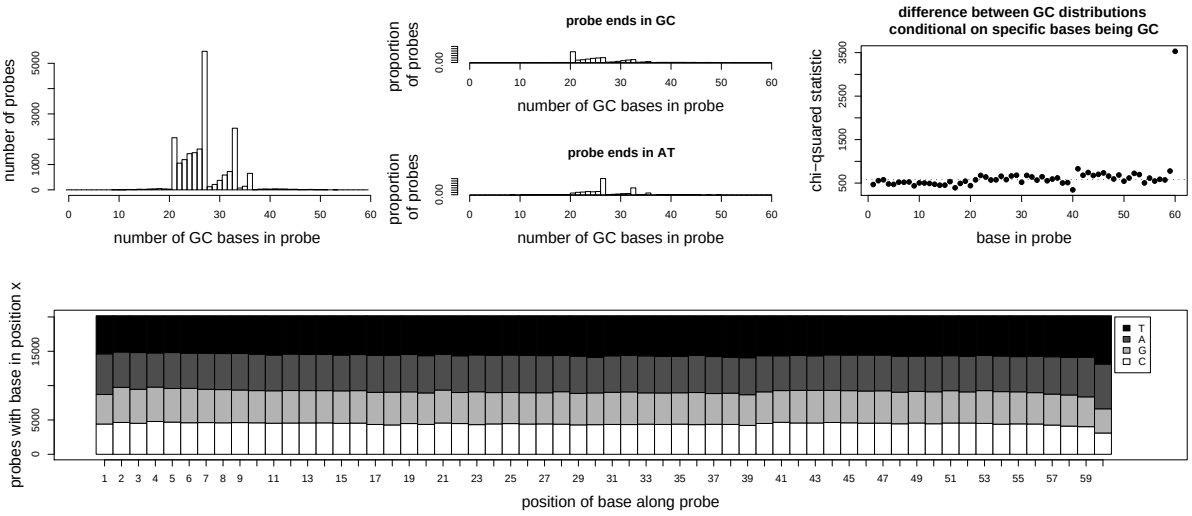

Figure 1: The structure of probes in the Agilent Human 1A Oligo Microarray.

Our data consist of 85 arrays on which either a renal cancer sample (75) or a normal control (10) has been compared to a common reference sample, the reference sample always being in the red channel. We also have access to a set of similar arrays where the red channel has been degraded by ozone contamination. Additionally we use two datasets from public repositories: 21 arrays from patients with type 2 diabetes described in Hayashi *et al.* (2006), and 29 microarrays from an experiment to detect the off-target gene-silencing effects of siRNAs as described in Birmingham *et al.* (2006).

3 Evidence and Modelling of bias

That a bias exists is quickly apparent. We see that the log-intensities in an array can vary with simple probe structure; in this case (Figure 2, left) higher for probes with more C bases. We see that the effects are not cancelled out within arrays; the log-ratios of intensities here (Figure 2, centre) decreasing with the number of A bases in the associated probe. We should note though that while the phenomenon of the log-intensities is seen in nearly all examples, that for the log-ratios can be far more variable. This variability though must contribute to the bias persisting between array comparisons, with the t-statistic for a comparison between cancerous and normal samples showing (Figure 2, right) a clear relationship with the number of G bases in the associated probes.

Also apparent is a certain heteroscedasticity, perhaps clearest in the central panel in Figure 2. Here variances are smaller for 'middling' numbers of A bases, and higher at the extremes. However other patterns are observed in other arrays.

It is possible, indeed almost certain, that the correct model will not be in terms of a single base count. Nevertheless while we may better model the bias with an alternative model, it can not be simply an artefact of our inadequate model. The heteroscedasticity on the other hand might well be. This is important as the consequences of the heteroscedasticity may be the greater (in terms of effects on gene lists), and yet the harder to remove.

In attempting to account for the bias, we consider models for the log-intensity in terms of the following quantities: 1) counts for the individual bases (A, C, G and T) which we denote B_C, 2) separate base effects at the 60 probe positions, denoted B_L, 3) counts for the 16 pos-

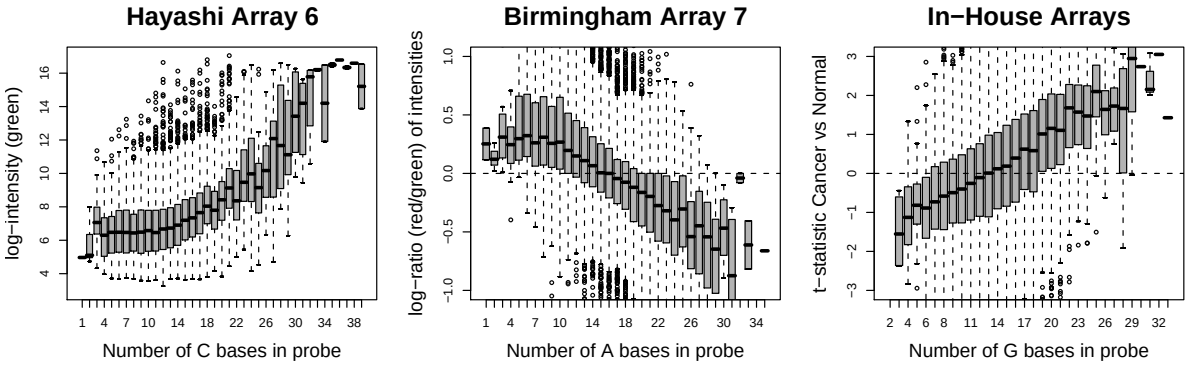

Figure 2: The persistent nature of bias due to probe design.

sible neighbouring pairs of bases, denoted P_C and 4) separate effects for the 16 pairs in the 59 possible locations, denoted P_L. The models we consider are $\log(I) \sim B_C$, $\log(I) \sim B_L$, $\log(I) \sim P_C$, $\log(I) \sim B_L + P_C$ and $\log(I) \sim P_L$ (where $\log(I)$ denotes log-intensity or log-ratio of intensities). Even this last model has a maximum of 886 parameters to estimate (in practice fewer), which is feasible given the number of observations on an array. The identity of the end base was also considered in models, but contributed nothing.

Assessment of model fits via the usual statistics is complicated since we are modelling the noise, and ignoring the effect (i.e. the varying intensities of genes), so we know that our models will not fit the data well in the usual sense. So as well as variance explained and AIC, we will use the $2,686$ probes with non-unique target gene as a validation set, fitting the model to the remaining $14,137$ probes and using the resulting model to correct the $2,686$; measuring whether probes hitting the same gene 'tighten up'. There are a number of other reasons why intensities, nominally from the same gene, might not agree (e.g. alternative splicing) so we neither expect nor desire to explain all the variance in this manner.

4 Results

As can be seen in Figure 3, the main point of interest is that while the models featuring only single-base information perform poorly in the single channels, they do a lot better in the log-ratios while the most complicated model shows abysmal performance in the log-ratios. Location

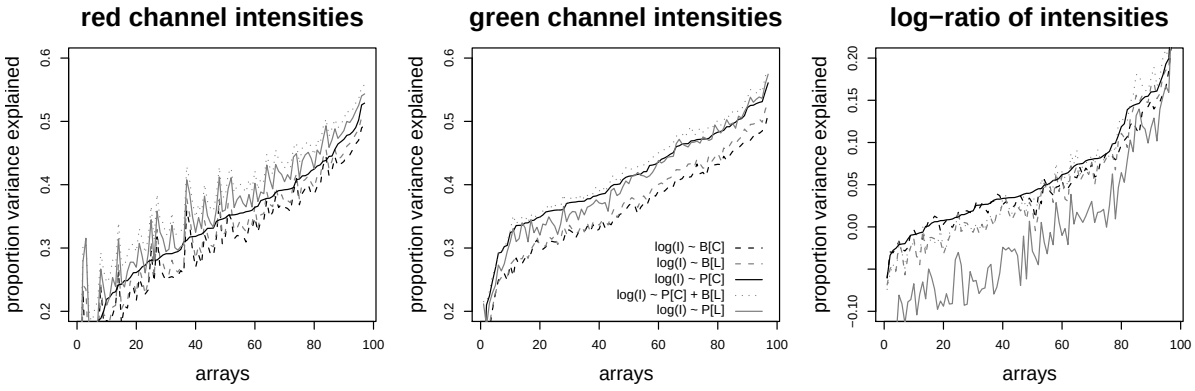

Figure 3: The performance of the models in explaining variance amongst the $2,686$ probes.

information looks to be of greater importance in the red channel and is of even greater importance in the arrays known to be ozone affected (not shown). In general modelling the log-ratios in terms of counts of neighbouring pairs of bases would appear to perform best.

5 Discussion

We surmise that the effect seen in the log-intensities is due to effects of the dye hindering the hybridisation of RNA to the probes. It is obvious then that the pattern in the log-ratios is driven by the discrepancy between the effects seen in the red and green log-intensities. That there should be such an effect in the results of a linear model applied across so many arrays we attribute to a combination of the heterogeneity of effects between arrays, the heteroscedasticity of results between probes, and the imbalance in the design of the experiment for comparing normals to cancers.

The effect is of obvious importance, both for the interpretation of published literature and conduct of future experiments. It has potential value as a method for quality control: Within log-ratios the magnitude of the effect varies between arrays and can be monitored and perhaps controlled. It has value in the critical appraisal of experiments: We can identify probes that were *a priori* more likely to be flagged as significant and perhaps a subset of trustworthy probes for a separate, robust, analysis. Finally we have characterized the effect in terms of the probe make-up in a manner that allows for a first attempt at correcting for the bias.

While the base-pair count model seems to offer the most value for correcting the effect, we stress that these are not simply the standard nearest-neighbour melting temperature values. Finally we note that the increased importance of base-location within models applied to the ozone affected data, is informative both regarding the ozone degradation process and potential diagnostics for it.

Acknowledgements

We thank Nuno Barbosa-Morais for the annotation and advice, and David Neal's group for access to some illustrative data.

References

Zhang L, Miles MF and Aldape KD (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, **21**:656-658

Abdueva D, Skvortsov D and Tavaré S (2006) Non-linear analysis of GeneChip arrays. *Nucleic Acids Research*, **34**:e105

Hayashi T, Urayama O, Kawai K, Hayashi K, Iwanaga S, Ohta M, Saito T and Murakami K (2006) Laughter regulates gene expression in patients with type 2 diabetes. *Psychotherapy and Psychosomatics* **75**:62-65.

Birmingham A, Anderson EM, Reynolds A, Ilsley-Tyree D, Leake D, Fedorov Y, Baskerville S, Maksimova E, Robinson K, Karpilow J, Marshall WS and Khvorova1 A (2006) 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods* **3**:199-204