

Spatial Defects in 5896 HG-U133A GeneChips

W. B. Langdon WLangdon@essex.ac.uk, R. da Silva Camargo and A. P. Harrison

Departments of Mathematical Sciences and Biological Sciences,
University of Essex, CO4 3SQ, UK

Abstract

Motivation: Modern biology has moved from a science of individual measurements to a science where data are collected on an industrial scale. Foremost amongst the new tools for biochemistry are chip arrays which, in one operation, measure hundreds of thousands or even millions of DNA sequences or RNA transcripts. Whilst this is impressive, increasingly sophisticated analysis tools have been required to convert gene array data into gene expression levels. Despite the assumption that noise levels are low, since the number of measurements for an individual gene is small, identifying which signals are affected by noise is a priority.

Results: 5896 raw data (Affymetrix CEL) files were obtained from <ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/TABM/E-TABM-185/>. Each CEL file was checked for spatial errors. In HG-U133A high-density oligonucleotide array (HDONAs) the mean error rate is only 1.6% which is amongst the best for human GeneChips. However some locations are much more error prone than others, with up to 28% of probes being affected.

Removal of erroneous data improves breast cancer survival prediction.

1 Introduction

There are increasingly large volumes of publicly available high-density oligonucleotide array (HDONAs) data. Chief amongst these are the EBI's ArrayExpress and NCBI's GEO. Almost 6000 Affymetrix Homo Sapiens HG-U133A GeneChips cel files have been collected from both and stored in ArrayExpress as experiment E-TABM-185.

Reimers and Weinstein (2005) found that when they introduced small amounts of spatial noise, many gene expression values more than doubled. Using RMA, in one case, they found that when two-fold noise was introduced to 5% of probes, 0.1% of genes appeared to change by more than 0.5 on a \log_2 scale. When they used MAS5, 3.4% of genes appeared to change. In most Affymetrix designs, large spatial errors affect between 2% and 5% of probes (Langdon *et al.* 2007). For E-TABM-185, its 1.6%, Table 1.

The bulk of the probe values follows an approximately log normal distribution, cf. Figures 3 and 4. However the tails are too heavy for a Gaussian. More than half of probes have one or more value outside six standard deviations. (We expect only 6, rather than more than 300 000). For an average probe, six standard deviations corresponds to more than a 50 fold change rather the 1.4 or 2 fold used by Reimers and Weinstein (2005).

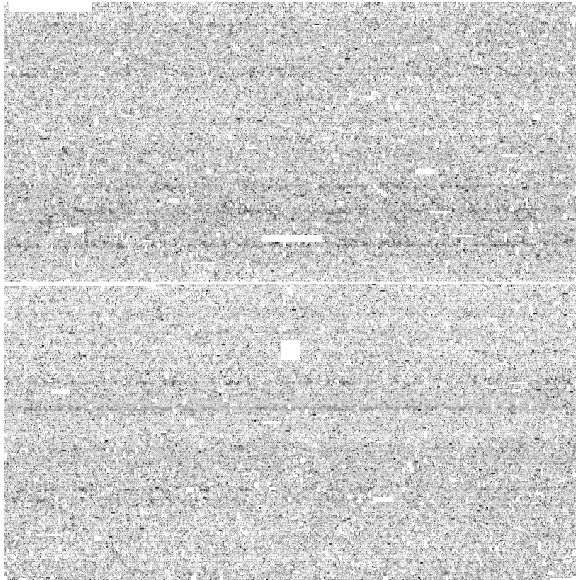


Figure 1: Average intensity 5896 HG-U133A arrays. To create a robust average, the top 0.5% and the smallest 0.5% of each probe's values are discarded, the geometric mean of the remaining intensities is calculated. Plotted average values lie between 75 and 27 078 (black). Controls are suppressed.

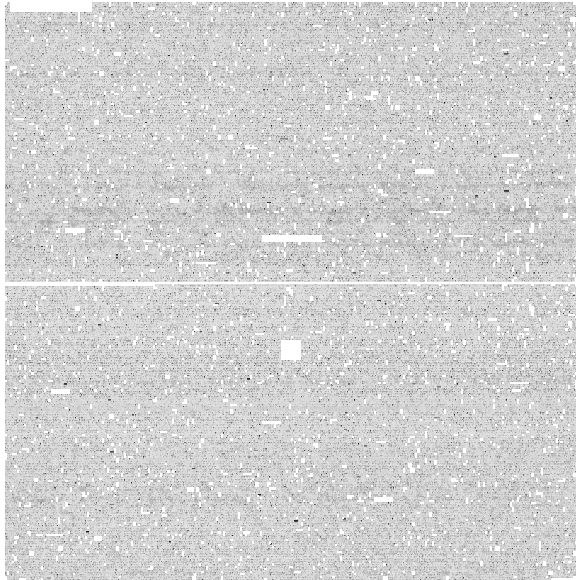


Figure 2: Variability of intensity 5896 HG-U133A arrays. I.e. standard deviation corresponding to mean of natural log values in Figure 1. Plotted (log) values lie between 0.4 and 2.0 (black) median 0.66 (i.e. 1.94 fold). (NB. discarding extreme values gives a small but systematic underestimate.)

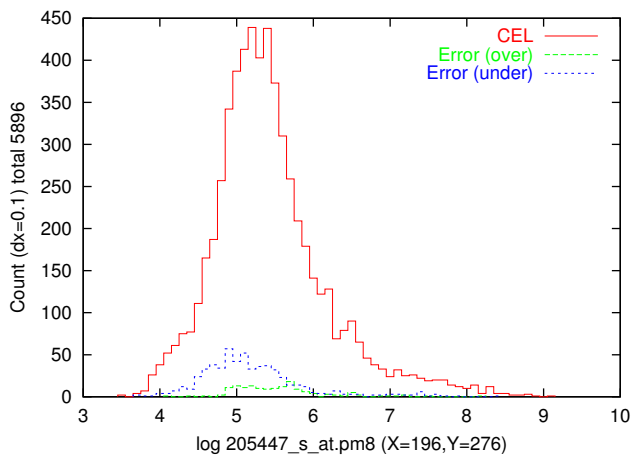


Figure 3: Distribution of intensities for an average probe. Two lower curves refer to probes within 4 of a spatial error.

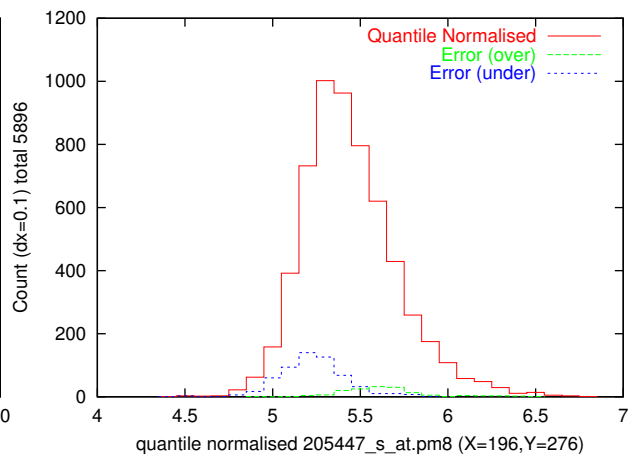


Figure 4: Distribution of probe (left) values after quantile normalising the whole chip to the average distribution (cf. Figure 1).

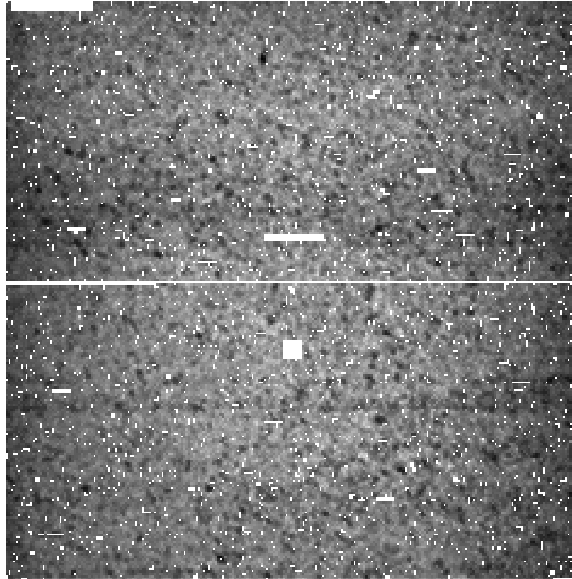


Figure 5: Location of spatial errors in 5846 HG-U133A arrays. Areas which are either too low or too high have been summed. In the worse case 28% (black) of cel files have erroneous probe values. Controls are suppressed.

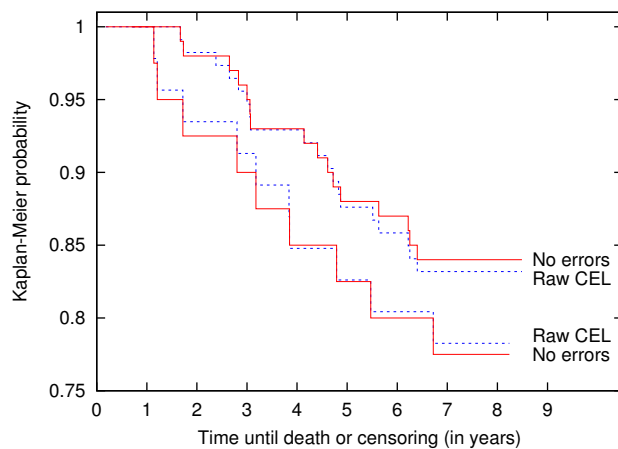


Figure 6: Improved Kaplan survival plots after removing errors (red). Survival predicted if $1.54 \frac{201893_x_at.2pm}{219260_s_at.7pm} - 2.94 \times 219260_s_at.7pm - \frac{219260_s_at.7pm}{200903_s_at.8mm} < 0$ on whole of Stockholm cohort (blue) and after removal of spatial errors effecting any of the three probes (red). Predictor created using genetic programming on an nVidia 8800 GTX GPU using 50 million data points (Langdon and Harrison 2008).

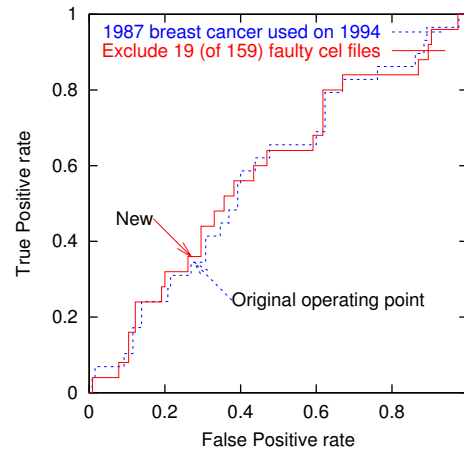


Figure 7: Comparison of two Receiver Operating Characteristics (ROC) curves showing removing flawed data improves non-linear combination of decorin, C17orf81, S-adenosylhomocysteine hydrolase prediction of breast cancer. The predictor trained on 49 E-TABM-185 cel files drawn from 1987-89 Uppsala cohort Miller *et al.* (2005) and tested on 159 patients from the 1994-96 Stockholm cohort Pawitan *et al.* (2005).

Table 1: Spatial flaws in E-TABM-185 Affymetrix GeneChip arrays. “miss” indicates those with missing data. “Reject” those with more than 10% spatial defects. To the left of the double bar, “Ok” gives the number of remaining cel files and the mean fraction of probes with spatial defects in the Ok files. This is split into those which are brighter than expected “over”. Less than average “under” and their sum “both”.

Chip	Number	miss	Reject	Ok	over	under	both
HG-U133A	5896	0	50 0.8%	5846	36 0.6%	60 1%	1.6%

We have automatically identified both bright and dim defects of widely varying sizes and shapes in many thousands of GeneChips of 13 different types (Langdon *et al.* 2007). All contained defects. See also Table 1.

2 Method

We updated Upton and Lloyd (2005)’s method to replace their requirement for technical replicants by comparison with an average chip. Essentially they assume in a perfect chip adjacent probes are statistically independent. Therefore spatial flaws can be found by searching for patches of the data which are consistently either above or below the average for that location. To avoid the possibility of correlations being introduced by contiguous perfect match (PM) and mismatch (MM) probe pairs, we follow Upton and Lloyd (2005) and use a checker board pattern to ensure we use only one of each pair. This diagonal pattern selects about half the probes to be compared with their average. As the data are normalised, the chance of any probe being above or below average is a half. The method is firmly grounded in statistics and simply highlights regions where more probes are brighter or dimmer than would be expected by statistical chance.

We calculate an average value for each probe by averaging the logarithms of the values observed at that location, thus calculating the geometric mean, cf. Figures 1 and 2. To avoid artefacts caused by outliers, we discard the upper and lower 0.5% of the observed values and calculate the geometric mean of the central 99%. This allows us to use the `affy` package in R in a single pass. Reimers and Weinstein (2005) also use a truncated mean on several 10s of spotted arrays as well as GeneChips.

As a second pass we compare all the arrays with the average to highlight suspiciously similar areas of each chip. Quantile normalisation *per row* is used to ensure each array has the same background reading. Using R on a modern PC it took more than five hours to process the 5896 HG-U133A cel files in E-TABM-185.

3 Results

In Table 1 we separate cel files with gross errors. The mean fraction of probe values affected by spatial errors, for the remaining chips, is given in the last column of Table 1.

There is a tendency for spatial errors to be associated with probe values being dimmer than expected and they tend to appear near the edges of chips, cf. Figure 5.

A breast cancer predictor had been previously created (Langdon and Harrison 2008). Its performance was improved on new data by excluding 19 cel files where there are spatial flaws within 4 units of the three probes it uses, cf. Figures 6 and 7.

4 Conclusions

It is known that even small defects in GeneChips reduce current analysis software’s ability to measure mRNA sequences (Song *et al.* 2007). Yet we find in published cel files, for a wide range of human tissues and disease conditions, errors in *all* HG-U133A data.

Our method is non-parametric and is based on firm statistical foundations rather than heuristics. Since it looks only at differences, it is insensitive to the strength of the noise. However Langdon *et al.* (2007) shows spatial noise can greatly exceed two-fold, either up or down. All it needs is the independence assumption and to be able to approximate the median with a truncated geometric mean. It does not assume the probe intensities follow a Gaussian, log normal, exponential or any other distribution. The technique is fully automatic¹. There is no need to manually adjust sensitivity or other tuning parameters.

Since newer high-density oligonucleotide array designs tend to have fewer measurements per genetic locus, we anticipate discovery and removal of spatial defects will become even more important in the future. For example, recent Affymetrix exon arrays have typically less than five measurements per genetic locus.

We suggest our automatic method is suitable for other species and other types of high-density oligonucleotide arrays. Previous studies (Reimers and Weinstein 2005) have shown that noise levels, much smaller than those we have observed in published data, seriously affect some gene expression levels calculated by common tools. We suggest flagging defects and removing suspect data will enable Biologists to safely use the remaining portions of expensive GeneChips.

Acknowledgement

I would like to thank G. J. G. Upton, Jose Manolo Arteaga-Salas and Lance Miller.

References

- 1 Langdon, W. B. and Harrison, A. P. (2008). GP on SPMD parallel graphics hardware for mega bioinformatics data mining. *Soft Computing*. Submitted to Special Issue 22 Sep 2007.
- 2 Langdon, W. B., Upton, G. J. G., da Silva Camargo, R., and Harrison, A. P. (2007). A survey of spatial defects in homo sapiens affymetrix genechips. In preparation.
- 3 Miller, L. D. *et al.* (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences*, **102**(38), 13550–5.
- 4 Pawitan, Y. *et al.* (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, **7**, R953–R964.
- 5 Reimers, M. and Weinstein, J. N. (2005). Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics*, **6**(166).
- 6 Song, J. S. *et al.* (2007). Microarray blob-defect removal improves array analysis. *Bioinformatics*.
- 7 Upton, G. J. G. and Lloyd, J. C. (2005). Oligonucleotide arrays: information from replication and spatial structure. *Bioinformatics*, **21**(22), 4162–4168.

¹ Recently Reimers and Weinstein (2005) and Song *et al.* (2007) have published interactive tools.