Row Quantile Normalisation of Microarrays

W. B. Langdon

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

Technical Report CES-484
ISSN: 1744-8050
23 June 2008

Abstract

Variation in tissue sample preparation leads to variation across the Transcriptome not just between experiments but to between individual microarrays. Normalisation is essential before data from different arrays can be compared. Quantile normalisation can be used to force data from a single GeneChip to take a given distribution. However quantile normalisation can be blind to the consistent spatial variation we note in thousands of Affymetrix' High-density oligonucleotide array (HDONAs) from NCBI GEO. We propose a simple computationally efficient normalisation technique which takes into account the spatial aspect. BioConductor R code is included.

Keyword: Normalisation, Gene expression, High-density oligonucleotide array, data mining, standard chip, MIAME

1 Introduction

While high-density oligonucleotide arrays have created biological data on an industrial scale, the cost of such GeneChips means few arrays are used in each experiment. Unfortunately it is difficult to control the dose of total mRNA. Instead some form of normalisation of the data collected by different arrays is necessary.

Early normalisation schemes tended to do things such as taking a number of readings from different samples and adjusting them to have the same mean. Typically the adjustment was made by linear rescaling. It was realised that by rescaling the data could also be adjusted to have the same standard deviation. In fact with modern computers it is possible to adjust data so that not only the first and second moments are the same but in fact the whole distribution is the same. This is known as quantile normalisation (Bolstad *et al.* 2003).

Typically quantile normalisation is applied across all the chips used in a particular experiment. However the advent of large public repositories of GeneChip experiments opens the way to normalisation against far larger numbers of chips. Normalisation against all published experiments using the same chip has been demonstrated (Langdon *et al.* 2007b). With a reference chip, created by averaging across all of GEO (Barrett *et al.* 2007), a Bioinformatician can quantile normalise a single chip. Further this can be done quickly using R.

Figure 1 shows 1 354 896 mRNA concentration measurements provided by a single Affymetrix HG-U133 2+ GeneChip. The horizontal spatial banding is obvious. (Figure 2 plots it numerically.) Figure 3 shows the banding carries over to individual probe performance. C and G DNA base pairings have three rather than two main hydrogen bonds. This means probes with many Cs or Gs bind more tightly than those with more As and Ts. (This is confirmed by Figure 4.) The number of each type of base in HG-U133 2+ probes follows a similar pattern to that which would be produced by a random choice, see Figure 6. The

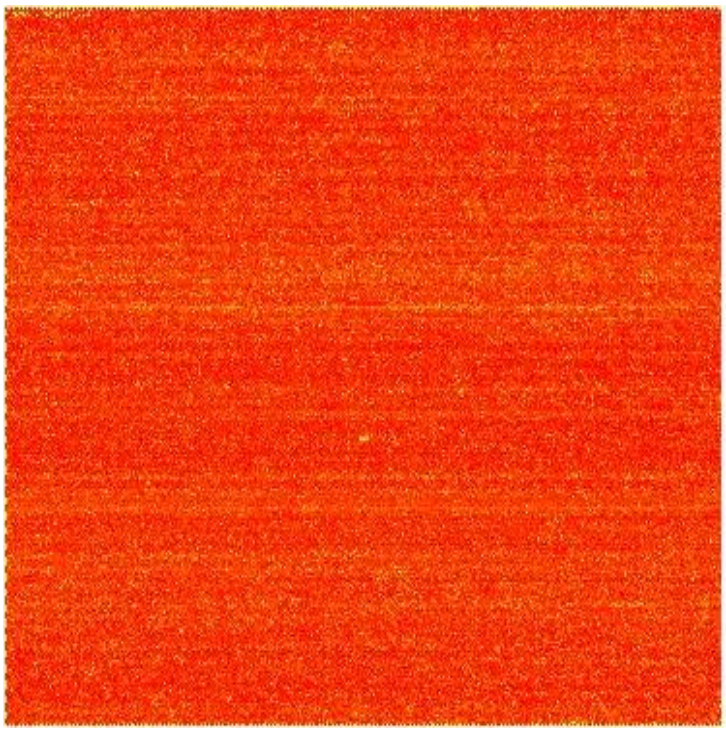

Figure 1: Example Homo Sapiens HG-U133 +2 GeneChip (prior to normalisation, yellow is higher). The banding of signal intensities is clear. Note perfect match (PM) probes are adjacent to their mismatch (MM) partner in the next row. However the horizontal banding occurs regardless of PM/MM. (In terms of spatial defects, this CEL file is the median of 2757 HG-U133 2+ down loaded from NCBI's GEO).

probes are deliberately placed on the chip such that probes with similar sequences are adjacent. This increases the feature size on the 100 photolithographic masks required for each GeneChip. Having larger holes in the masks, caused by adjacent probes have the same base at the same depth, tends to reduce the importance of edge effects. This in turn eases manufacture (Pease *et al.* 2001; Feldman and Pevzner 1994). Assorting the probes by sequence gives the non-random layout of the probes shown in Figure 5. It is this assorting that is the underlying cause of the banding in both probe intensity and performance. Until now this strong spatial feature has been ignored by common normalisation tools.

2 Row Quantile Normalisation

Recently we have applied quantile normalisation not to just a collection of GeneChips relating to one experiment but to thousands of public chips (Langdon *et al.* 2007a). This allows a Biologist to normalise a single chip against a public base line (i.e. GEO), rather than his own much smaller private experimental set. And normalisation can be safely done for a small number of chips, even just one.

Quantile normalisation works by placing measured data in rank order. The reference distribution (in our case a robust average taken across thousands of GeneChips of the same type in GEO) is also sorted into rank order. To normalise a GeneChip, the sorted probe intensities are simply replaced by the corresponding value with the same rank from the reference distribution, i.e. the smallest measured probe intensity is replaced with the smallest average value in GEO. The value of the probe with second smallest value is replaced with the second smallest average value in GEO. And so on, until the largest measure probe intensity has been replaced by the largest average probe value from GEO. This can be done efficiently in R. Note with ordinary quantile normalisation a probe can be normalised to the average value of any probe. That probe may be anywhere on the GeneChip.

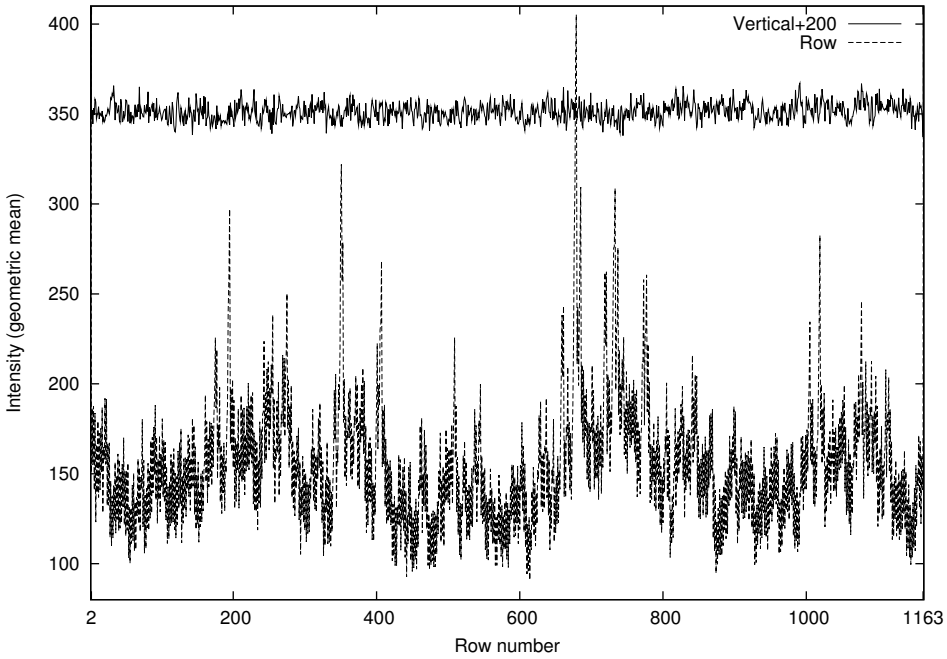

Figure 2: Average intensity by row (dashed line). In contrast average signal across rows is close to 150. Variation is about a factor of two, both along rows and across them. (Same data as Figure 1.)

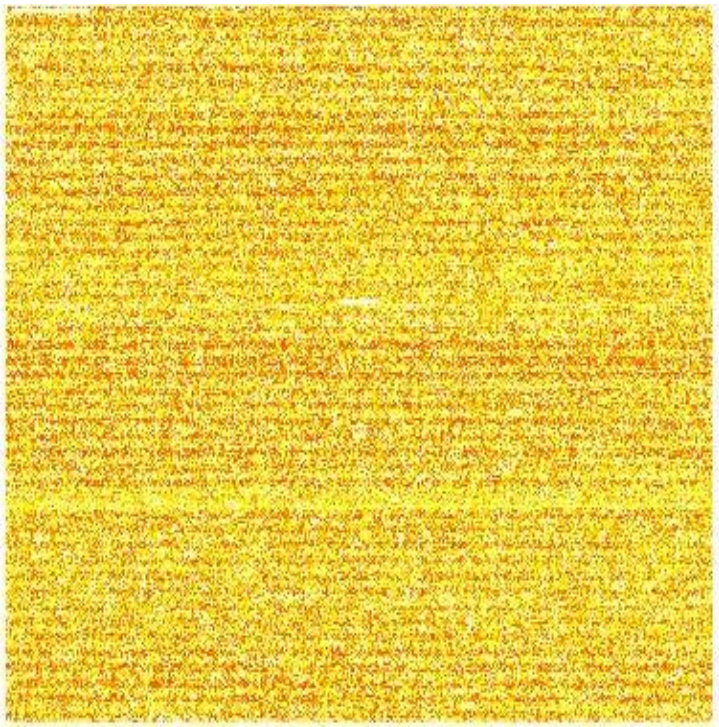

Figure 3: Performance of probes across HG-U133 +2 GeneChips. Highest correlation between probes in the same probeset. (2757 GEO HG-U133 +2). (Yellow is higher. Areas without probesets are white.). The banding of signal performance is clear.

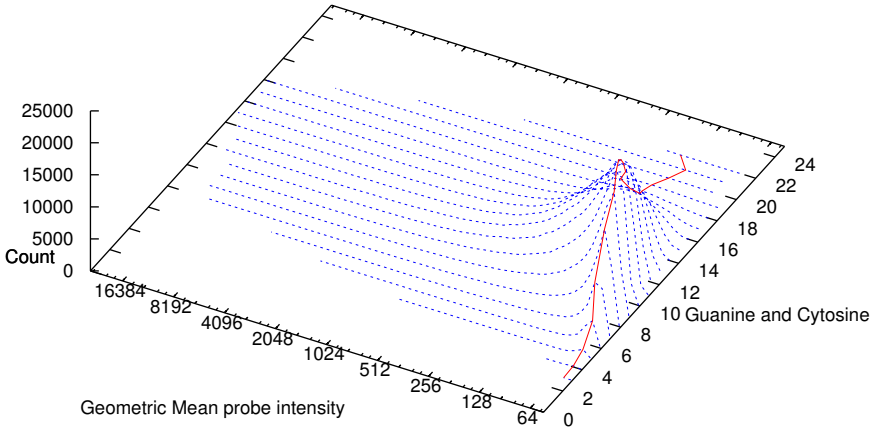

Figure 4: Distributions of average HG-U133 2+ probe intensity vs. number of Cs and Gs. On average (mode, solid line) probes with many C+G are approximately twice as bright as those with few.

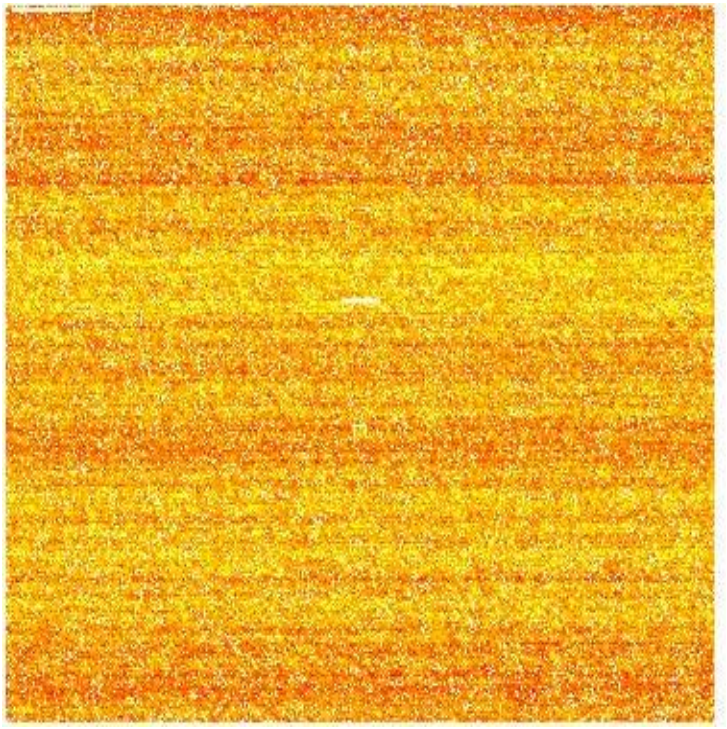

Figure 5: Number of C's and G's in probes across HG-U133 +2 GeneChips. (Yellow is higher. Areas without probesets are white.).

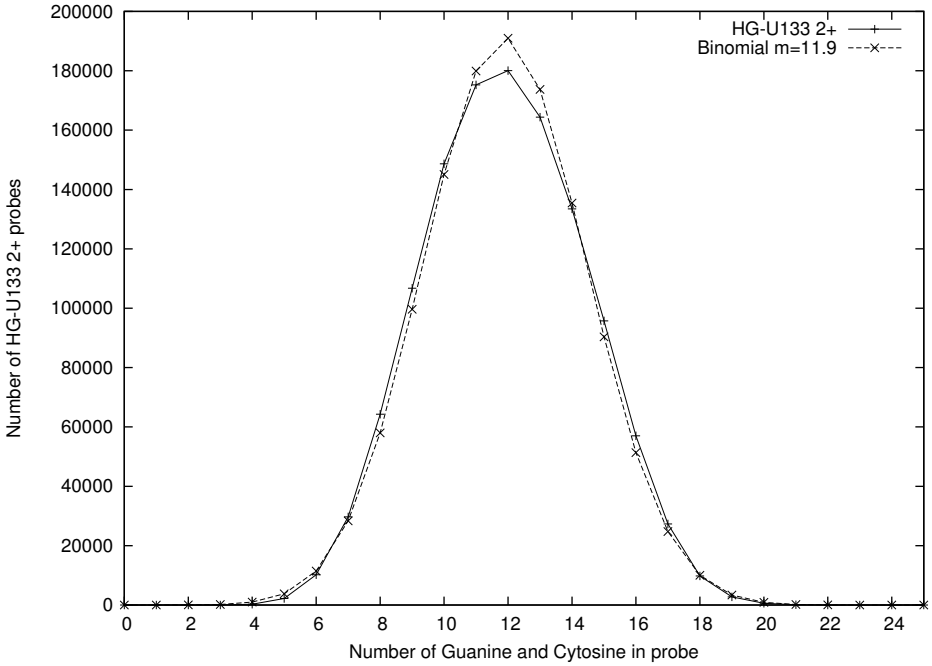

Figure 6: The distribution of number of C's and G's in HG-U133 2+ probes is approximated by a random distribution (dashed line) with the same mean (11.9). (Same data as Figure 5.)

Row quantile normalisation is the same as regular quantile normalisation, except we insist that a probe's normalised value must come from a probe in the same row. With modern GeneChips, this still leaves the sort algorithm hundreds or thousands of values to chose from. Using the same row automatically takes advantage of the way Affymetrix lays out probes on its GeneChips (see previous section). This ensures every probe in normalised to the average value of a probe which has a similar probe sequence to it.

Row normalisation is readily and efficiently implemented in R. See code fragments in Figure 7.

References

1 Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, **35**(Database issue), D760–D765.

2 Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

3 Feldman, W. and Pevzner, P. A. (1994). Gray code masks for sequencing by hybridization. *Genomics*, **23**, 233–235.

4 Langdon, W. B., da Silva Camargo, R., and Harrison, A. P. (2007a). Spatial defects in 5896 HG-U133A GeneChips. In J. Dopazo, A. Conesa, F. Al Shahrour, and D. Montener, editors, *Critical Assesment of Microarray Data*, Valencia. Presented at EMERALD Workshop.

5 Langdon, W. B., Upton, G. J. G., da Silva Camargo, R., and Harrison, A. P. (2007b). A survey of spatial defects in Homo Sapiens Affymetrix GeneChips. Submitted.

6 Pease, R. F., McGall, G., Goldberg, M. J., Rava, R. P., Fodor, S. P. A., Goss, V., Stryer, L., and Winkler, J. L. (2001). Printing molecular library arrays. US Patent 6,239,273. Affymetrix, Inc. (Santa Clara, CA), Filed: October 26, 1999.

```
sortbyrow=function(mat) {
  apply(mat,2,sort);
}
quantilebyrow=function(mat,rank) {
#rather than sort, use rank to ensure two probes with the same value
#before normalisation, have the same value after normalisation.
  r = apply(mat,2,rank,ties.method="min");
  ans=array(NA,dim=c(nrow(rank),ncol(rank)));
  for(j in seq(from=1,to=ncol(r))) { ans[,j]=rank[r[,j],j]; }
  return(ans);
}



qdistribution <- sortbyrow(normean);
library(affy);
a   <- ReadAffy(filenames,sd=FALSE);
mat <- quantilebyrow(array(exprs(a),dim=c(S,S)),qdistribution);
```

Figure 7: R code fragment showing the row quantile normalisation of a single GeneChip `filenames` against the whole of GEO. The reference distribution, previously calculated from all the GeneChips in GEO of the same type, is stored in `S`×`S` array `normean`. (For an HG-U133 2+ `S` = 1164.) The R function `sortbyrow` orders the array's rows independently. The individual probe values, given by `ReadAffy` and `exprs(a)`, must also be stored in a `S`×`S` array. `quantilebyrow` creates an array of indexes `r` into the reference distribution `rank`. For each row, it calculates a vector of `S` values which are taken from `rank` via the sorted indexes in `r`. (The R `[,j]` syntax implies all elements of an array with second index `j`.)