Probe Level Quantile Normalization of High Density Oligonucleotide Array Data

Ben Bolstad
Division of Biostatistics, University of California, Berkeley

December 2001

Introduction

To reliably compare data from multiple chips one needs to minimize non biological differences that may exist. One process that helps is to normalize within a set of chips. We will propose a method that can quickly normalize within a set of chips without choosing either a baseline chip to which all chips are normalized or working in a pairwise manner. The method will deal reliably with non linearities.

Method assumption

The method assumes that there is an underlying common distribution of intensities across chips. See figure 1 for an example.

Motivation for the algorithm

One can use a qqplot as a tool to compare if two datasets come from the same distribution. If they are from the same distribution then the quantiles line up on the diagonal. See figure 2

This suggests that one could give two disparate datasets the same distribution by transforming the quantiles of each to have the same value. This could be done by projecting onto the unit diagonal $\left(1/\sqrt{2}, 1/\sqrt{2}\right)$. Extending this idea to N dimensions gives us a method of finding a common distribution from multiple data vectors.

Algorithm

1. Given N datasets of length p form X of dimension $p \times N$ where each dataset is a column
2. Set $d = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)$
3. Sort each column of X to give X_{sort}
4. Project each row of X_{sort} onto d to get X'_{sort}
5. Get X_{norm} by rearranging each column of X'_{sort} to have the same ordering as original X

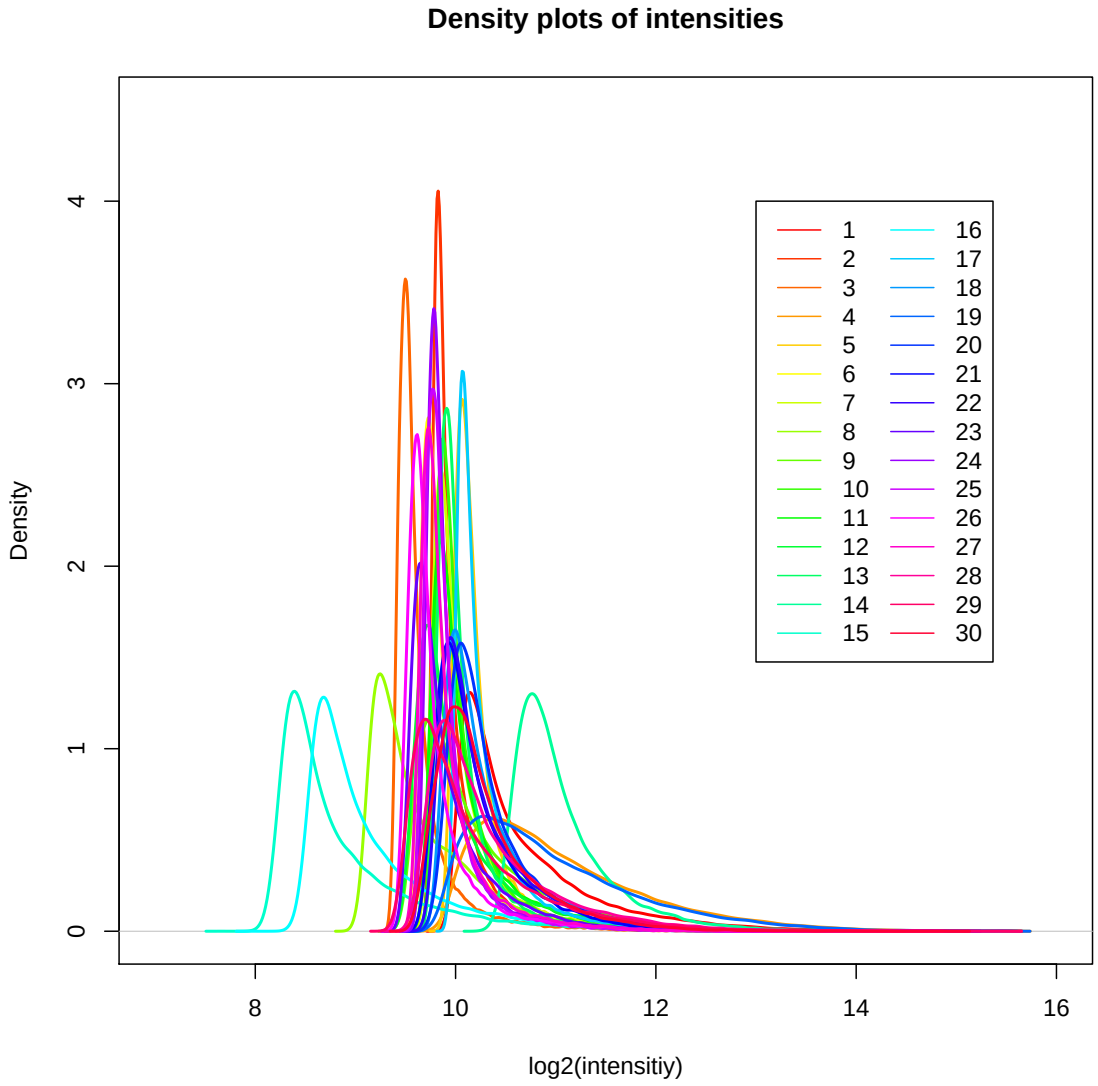

Figure 1: Density plots of cell intensities across chips

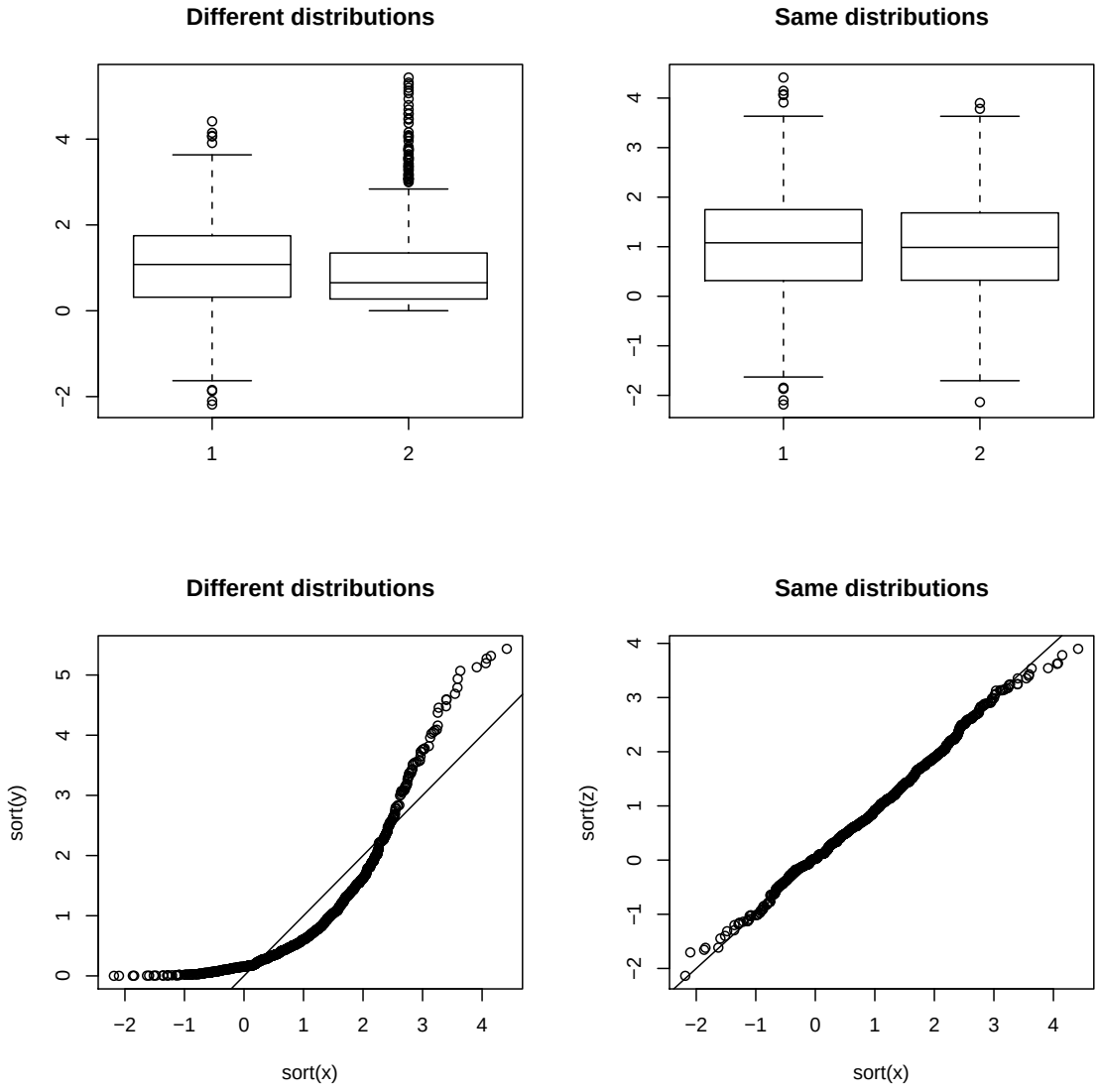

Figure 2: QQ plots

Notes

1. If $q_i = (q_{i1}, \ldots, q_{iN})$ is a row in X_{sort} then the corresponding row in X'_{sort} is given by $\boldsymbol{q}'_i = \text{proj}_{\boldsymbol{d}}\boldsymbol{q}_i$
2. The projection is equivalent to talking the average of the quantile in a particular row and substituting this value for each of the individual elements in that row

$$\text{proj}_{\boldsymbol{d}}\boldsymbol{q}_i = \frac{\boldsymbol{q}_i \cdot \boldsymbol{d}}{\boldsymbol{d} \cdot \boldsymbol{d}}\boldsymbol{d} = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} q_{ij}\boldsymbol{d} = \left(\frac{1}{N}\sum_{j=1}^{N} q_{ij}, \ldots, \frac{1}{N}\sum_{j=1}^{N} q_{ij}\right)$$

An example

We treat both PM and MM values as intensities to be normalized together. When using only PM's to construct your expression measure one might consider only PM values and normalize them while ignoring the MM values. The original densities of PM and MM intensities by chip with the normalized distribution from applying the method to PM and MM intensities superimposed in black as shown in figure 3. Also we have boxplots of intensities pre and post normalization as shown in figure 4. Finally you can see a selection of pairwise PM plots shown first before normalization in figure 5 and then after quantile normalization in figure 6. The blue lines are lowess smoothers. So we can see that the normalization was successful.

Possible problems

One problem with this method is that in the tails in particular, where we might expect greater differentiation between chips, the normalized values are going to be identical. A modification has been implemented that allows greater differentiation. This works by scaling and centering extreme tail values appropriately without affecting the corresponding quantiles in the other chips. Boxplots can be seen in figure 7. Other modifications are also being explored.

Generally speaking one is computing an expression measure for a probeset based upon multiple PM probes or PM/MM probe pairs, thus it may be acceptable to use the uncorrected method.

For smaller numbers of chips, especially when dealing with just 2 chips, a pairwise normalizer may be preferable.

Conclusions

The quantile based method provides a fast method to normalize multiple chips, provided one is willing to assume a common distribution.

Acknowledgments

I would like to thank Terry Speed and Rafael Irizarry for useful comment and criticism.

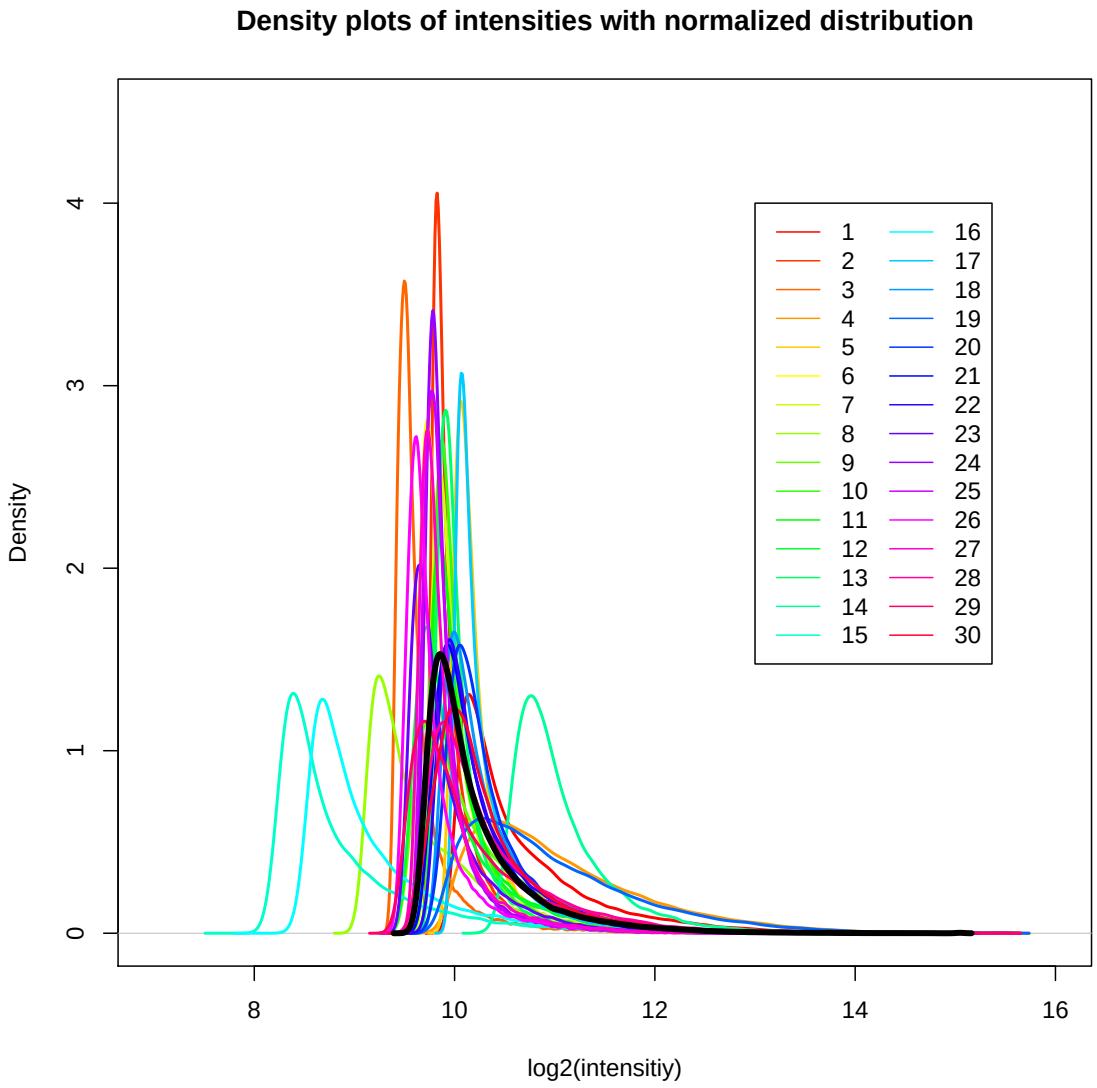

Figure 3: Cell intensity distributions with the normalized distribution

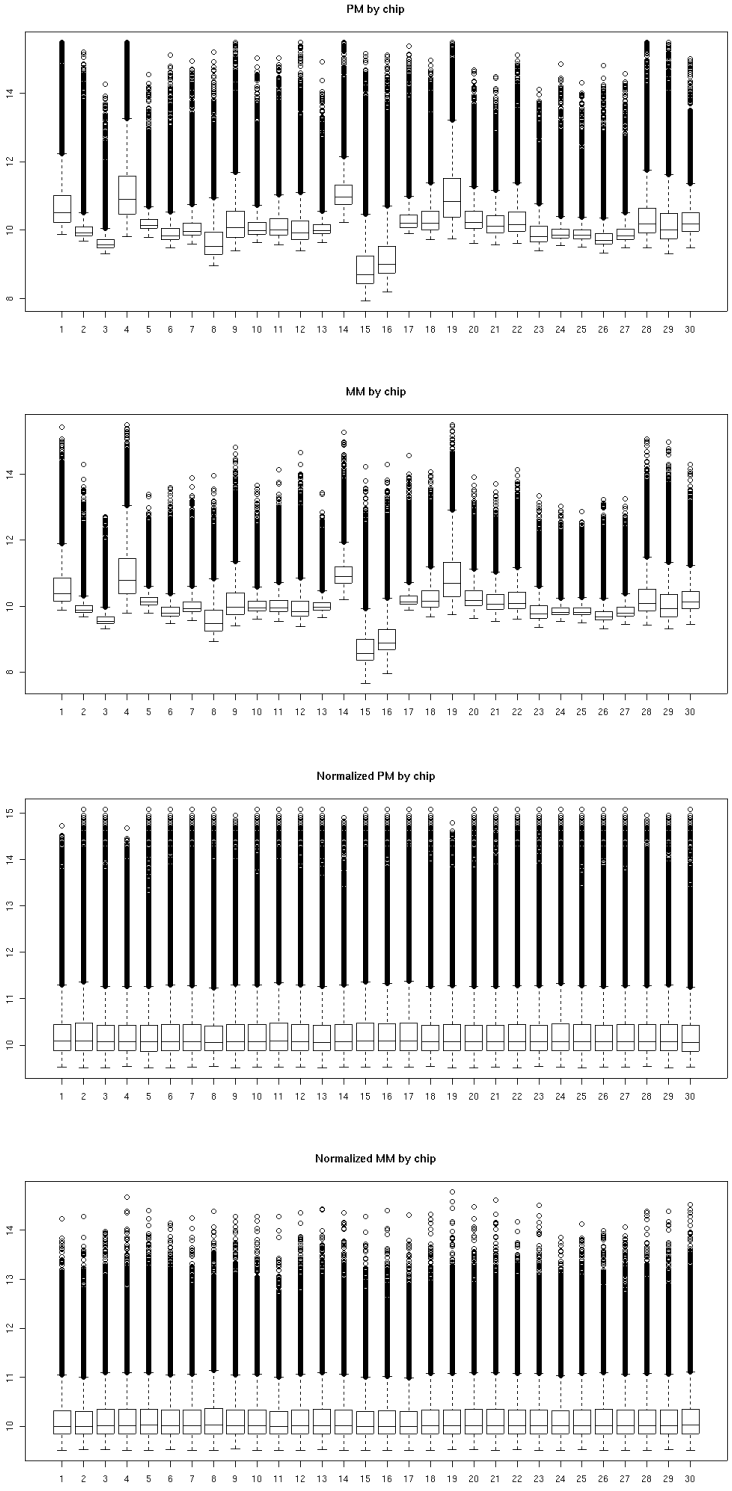

Figure 4: Before and after normalization boxplots of intensities by chip.

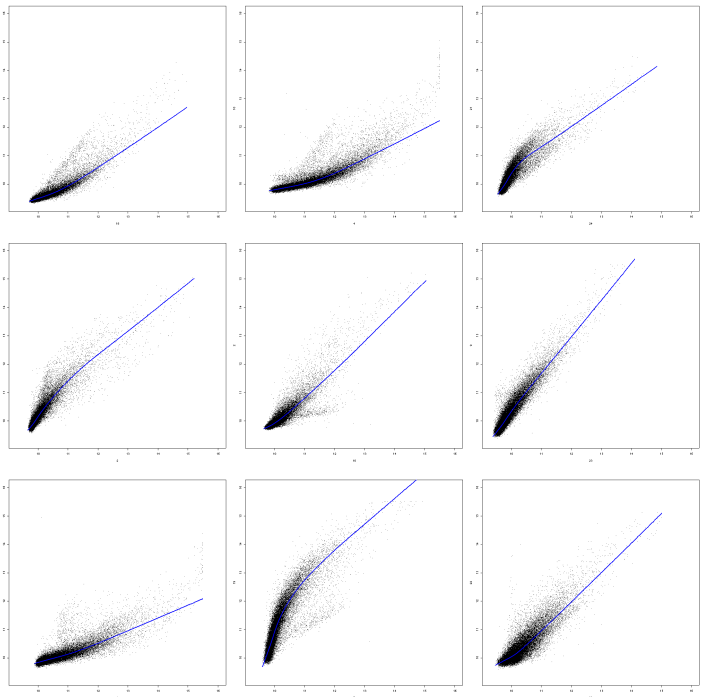

Figure 5: Before normalization pairwise PM

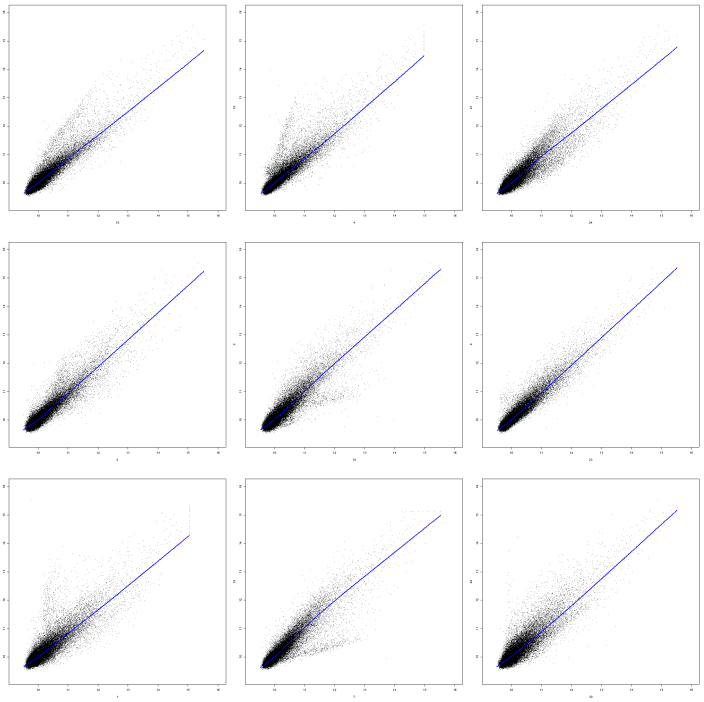

Figure 6: After normalization pairwise PM.

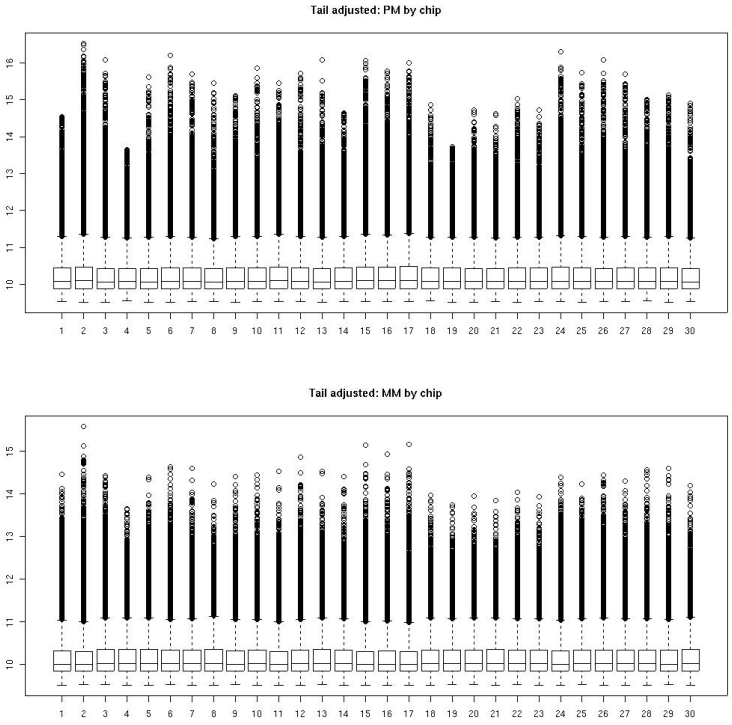

Figure 7: Tail adjustment normalization