



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL of PUBLIC HEALTH

---

Johns Hopkins University, Dept. of Biostatistics Working Papers

---

6-26-2009

# Subset Quantile Normalization using Negative Control Features

Zhijin Wu

*Brown University, [zwu@stat.brown.edu](mailto:zwu@stat.brown.edu)*

---

## Suggested Citation

Wu, Zhijin, "Subset Quantile Normalization using Negative Control Features" (June 2009). *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 191.  
<http://biostats.bepress.com/jhubiostat/paper191>

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

Copyright © 2011 by the authors

# Subset Quantile Normalization using Negative Control Features

Zhijin Wu\*

Center for Statistical Sciences and Department of Community Health, Brown University

Email: zhijin\_wu@brown.edu;

\*Corresponding author

## Abstract

---

Normalization has been recognized as a necessary preprocessing step in a variety of high throughput biotechnologies. A number of normalization methods have been developed specifically for microarrays, some general and others tailored for certain experimental designs. All methods rely on assumptions on what values are expected to stay constant across samples. Most assume some quantities related to the specific signal stay the same, which is usually not verifiable. Some recent platforms has began to include a large number of negative control probes that cover a great range of the measured intensity. Using these probes as normalization basis finally makes it possible to normalize without making assumptions on the behavior of biological signal. We present a subset quantile normalization procedure that normalizes based on the distributions of nonspecific control features, without restriction on the behavior of specific signals. This method is compared to two other leading non-linear normalizations, quantile normalization and loess normalization, using data from Affymetrix Tissue experiment on Human Gene 1.0 St Array. This method preserves the most biological variation after normalization while reducing the noise observed on control features. Although the illustration dataset is from an gene expression experiment, this method is general for all platforms that include a large set of control features, regardless of the biomedical application. It does not require equal number of features in all samples and tolerates missing data.

---

## 1 Introduction

High throughput biotechnologies have become increasingly important in biomedical research. Among these, DNA microarrays is probably the most widely used in applications studying variations of the transcriptome, the genome and the epigenome. The microarray technology simultaneously quantifies a large number of DNA or RNA species with various sequences by labeling the target sample with fluorescent dyes and hybridizing it to features (probes) with complimentary sequences of the target molecules. The target concentration is reflected in the fluorescent intensities observed on the complementary feature. Because the hybridization efficiency of each feature is different and unknown, microarrays only provide a measure of relative abundance of the target molecules. Thus the the quantity of interest on each feature is the biological variation of its target molecule in different samples. In addition to the quantity of interest, a number of other factors affect the feature intensities on an array in a systematic way. These factors include sample preparation, hybridization and array processing. The systematic variations caused by these factors are of no biological interest and are sometimes referred to as “*obscuring variations*” (Bolstad *et al.*, 2003). Sometimes the obscuring variations even exceed the level of biological variations. It has been well recognized that normalization is a necessary step to remove or reduce such variations and make the data from different arrays more comparable before further analysis is carried out.

A number of normalization methods have been developed since the introduction of the microarray technology. Some methods are general and some are tailored towards specific experimental designs. Regardless of the biological applications, all normalization methods reflect the observation that many factors causing obscuring variations affect the entire array in some systematic fashion, otherwise the biological and technical variations on each feature would not be identifiable. For example, one sample may have higher labeling efficiency and the feature intensities in this sample tend to be higher in general than that from other samples, or one scanner may give higher readings than another scanner. If the overall hybridization in different samples is not expected to differ, we would like to remove the global “array effect” in normalization. When the labeling or scanner effect is the same for all features, a scaling normalization such as aligning the medians or means of each sample would suffice.

This example illustrates a key issue in normalization: what is expected to be constant across samples? All normalization methods make such assumptions, explicitly or implicitly. Normalization is then achieved by equalizing certain summary statistics based on the assumption. The scaling normalization works well only when the obscuring variation is a linear effect, which is rare in reality. A number of non-linear methods have been proposed to allow more flexible normalization. Using gene expression as an example, if one

assumes that only a small set of genes have differential expression, or that up- and down- regulation are approximately symmetrical, the distribution of gene expression measures on an array should be the similar across all samples. *Quantile normalization* (QN) (Amaratunga and Cabrera, 2001; Bolstad *et al.*, 2003) can be applied in such cases and has been demonstrated to have favorable properties (Bolstad *et al.*, 2003). Loess normalization removes intensity-dependent biases in differential expression and works well under similar assumptions (Yang *et al.*, 2002) Sometimes a group of house keeping genes are assumed to have constant expression across samples and are used as internal controls for normalization. However, there have been numerous reports that housekeeping genes are found to be quite variable in given situations (Thellin *et al.*, 1999). Another choice is to use a rank-invariant set of genes (Tseng *et al.*, 2001), whose expression levels remain a similar rank on an array across samples, and normalize so that expression measures of these genes are constant. The size of the rank-invariant set depends on the sample data and measurements from this set may not span the entire intensity range (Yang *et al.*, 2002).

All of the normalization methods mentioned above have one aspect in common: assumptions are made on the stability of expression levels of certain genes, that is, assumptions on the behavior of specific biological signal. These assumptions are rarely verifiable. Recently, a series of oligonucleotide arrays with a new design became available, including the Affymetrix Exon arrays, Gene St arrays, tiling arrays and Illumina arrays. These arrays include a good number (over ten thousand) of negative control probes that are designed to monitor the extent of nonspecific binding and to assist background estimation and correction. Consistent with previous observation (Wu *et al.*, 2004), although designed to measure background noise, the intensities observed on these negative control probes cover almost the entire range of intensities from all features, as shown in Figure 1.

The availability of a large number of negative control probes makes it possible to observe the impact of systematic obscuring variations that normalization procedures hope to remove. These probes are designed to have no complementary match to the genome/transcriptome of the targeting species, thus we expect their specific binding induced intensities to be constantly zero. The nonspecific binding intensities, however, would be affected by the factors that have overall impact for the entire array, such as labeling efficiency, scanner setting, time of experiment etc. This makes them the perfect controls for normalization purpose, since we do expect the intensities on these features to stay constant regardless of how the biological signal varies from sample to sample. For the first time, we do not have to make assumptions on the specific biological signals to do normalization.

In this paper, we propose a normalization methods based on a group of negative control features that are

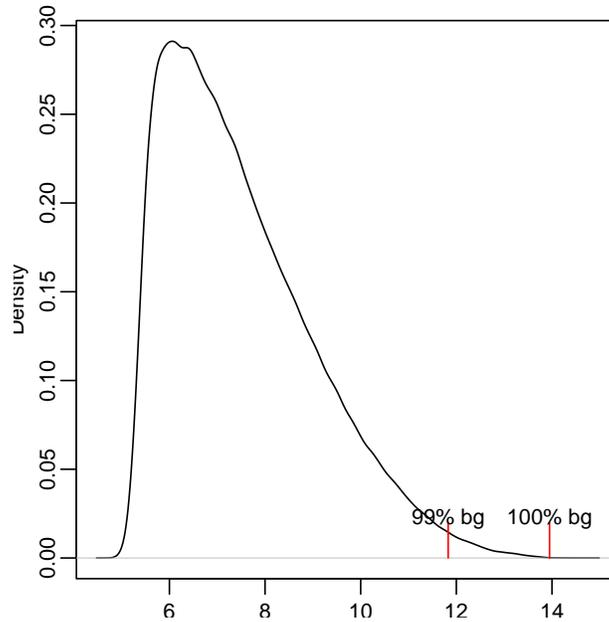


Figure 1: Probability density function of feature intensities ( $\log_2$  scale) from an Affymetrix Human Gene 1.0 St array. The 99<sup>th</sup> percentile and the maximum of negative control (bg) probes are marked to show that the intensity of control probes span almost the entire range of probe intensity.

designed to measure nonspecific binding, without restriction on the biological signal. We present the results on an gene expression experiment and demonstrate the improvement in preserving biological variation while reducing systematic noise. However, the application of this normalization method is applicable to all platforms that include such controls, regardless of the technology or biomedical application.

## 2 Data

For illustration, we use the Affymetrix *Tissue* experiment dataset on Human Gene 1.0 ST Array. A collection of 11 tissues (brain, breast, heart, kidney, liver, pancreas, prostate, skeletal muscle, spleen, testes and thyroid), each with three biological replicates, are included in the experiment, giving a dataset of 33 samples (arrays). This data set is provided to the public by Affymetrix at [http://www.affymetrix.com/support/technical/sample\\_data/gene\\_1.0\\_array\\_data.affx](http://www.affymetrix.com/support/technical/sample_data/gene_1.0_array_data.affx) On this platform, there are 16,943 negative control probes and 764,885 perfectMatch probes.

### 3 Methods

The proposed normalization method does not make assumption on the property of biological signals in different samples. Instead, we use quantiles of the negative control probes as “anchors” and require that these statistics are equalized after normalization. Intensities of all probes on an array are adjusted according to their relationship to the control quantiles on the same array. We term this method the “subset quantile normalization”, in order to differentiate with the complete QN that makes the distributions of the entire array equal.

Specifically, for each array  $a$ , we estimate the cumulative distribution function (CDF)  $F_a$  of the subset of probes that serve of controls. We use the estimates of  $F_a$  to define a reference control distribution  $F$  as the target distribution for the control probes. Now consider any probe on an array, if its raw intensity equals the  $q^{th}$  quantile of the control probes on the same array, its normalized intensity is defined as the  $q^{th}$  quantile from the reference distribution  $F$ . That is,

$$\tilde{y}_{a,j} = F^{-1}\{F_a(y_{a,j})\}.$$

We can estimate  $F_a$  by the empirical CDF  $\hat{F}_a$  and use the median of each quantile to define a reference control distribution  $\hat{F}$ . This works well for most of the data except the tails of the distributions because  $\hat{F}^{-1}$  is bounded by the observed intensities from the control probes. The probes whose intensity is beyond the maximum of the control set would have normalized values truncated at  $\hat{F}^{-1}(1)$ . To avoid the problem in tail areas, we use a semi-parametric approach. We first estimate  $F_a$  parametrically as a mixture of normal distributions,

$$\Phi_a^k(x) = \sum_{i=1}^k \pi_{ai} \Phi\left(\frac{x - \mu_{ai}}{\sigma_{ai}}\right),$$

where  $\Phi$  is the standard normal CDF.

The final estimate of  $F_a$  is defined as a weighted average of the empirical CDF and the normal mixture CDF,

$$\tilde{F}_a(x) = w\Phi_a^k(x) + (1 - w)\hat{F}_a(x).$$

$F$  is defined with a similar approach. From each array  $a$ , let  $y_{a,(k)}$  be the  $k^{th}$  order statistic of the control probes. The medians of the order statistics,  $\bar{y}_{a,(k)} = \text{Median}_k(y_{a,(k)})$ , define the reference control intensity vector. We use the values  $\bar{y}_{a,(k)}$  to estimate a normal mixture distribution  $\Phi^k = \sum_{i=1}^k \pi_i \Phi\left(\frac{x - \mu_i}{\sigma_i}\right)$  and compute a weighted average of  $\Phi^k$  and the empirical CDF,

$$\tilde{F}(x) = w\Phi^k(x) + (1 - w)\hat{F}(x).$$

With the CDF estimates  $\tilde{F}(x)$  and  $\tilde{F}_a(x)$  available, we define the normalized intensities as

$$\tilde{y}_{a,j} = \tilde{F}^{-1}\{\tilde{F}_a(y_{a,j})\}.$$

We leave  $k$  and  $w$  as tuning parameters for smoothness. Given  $k$ , the parameters for the mixture distributions  $\pi_s, \mu_s$  and  $\sigma_s$  can be estimated by maximum likelihood using EM algorithm. In practice,  $k = 5$  gives very good flexibility to a wide variety of distributions. We have used  $k = 5$  and  $w = .9$  in our example. Our implementation of the method uses the R package *Mclust* (Fraley and Raftery, 2009) for the EM estimation of mixture parameters and the R package *nor1mix* (Mchler, 2007) for the computation of mixture distribution quantiles.

## 4 Results

We compare the proposed normalization, subset QN, with two other most widely used nonlinear normalization methods, the complete QN, and the loess normalization, both found to perform well and are incorporated in numerous preprocessing modules (Smyth, 2005; Yang *et al.*, 2007; Gautier *et al.*, 2004).

There are several variants of the loess normalization and we have used the cyclic loess normalization (Bolstad *et al.*, 2003) in this comparison.

First, we examine the distributions of raw intensities on the negative control probes. Since the sequences of these probes are not present in the human genome, we expect their intensities to reflect the systematic array variation instead of biological variation. The empirical distributions of the raw intensities on these antigenomic probes appear to have different location and scale, as demonstrated by the various medians and inter-quartile-ranges in Figure 2A, indicating the need for normalization.

The loess and complete QN normalization do not make a lot of progress in making these control probes more comparable across samples, as seen in Figure 2A and 2B. This shows that after loess or complete QN, what is expected to be constant may still have considerable variation. In contrast, the subset quantile normalization forces the control probes to have the identical distribution by design (Figure 2D). The loess normalization is very similar to the complete QN in all comparisons to follow. The results from loess normalization are therefore omitted in the main text and are provided in supplementary material.

The goal of normalization is certainly not just to make the control data stable. More importantly, we want to reduce or remove the obscuring variability on the signal-bearing probes. Since the variation observed in the raw data is a combination of biological variation and obscuring processing variation, we would expect a good normalization procedure to reduce the variation among replicates and make them more similar. We

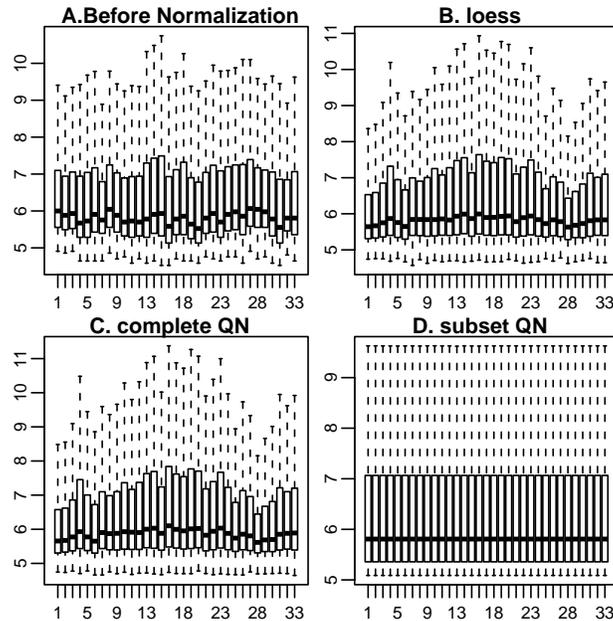


Figure 2: The boxplots of the feature intensities ( $\log_2$  transformed) from the negative control probes in the 33 samples from 11 tissue types. A. Raw intensities without normalization. B. Cyclic loess normalization is applied to the complete set of probes including the perfectMatch probes and the negative controls. Shown are boxplots of the negative controls after normalization. C. Like B but with complete quantile normalization. D. Like B but with subset quantile normalization.

thus compute the within-tissue variances for all 11 tissue types and average the 11 variances for each probe as a pooled within-tissue variance. Since the probe intensity variance is commonly observed to vary across log intensity levels, we stratify the probes into 20 groups based on the average raw intensity level.

Figure 3 compares the pooled within-tissue variances before and after normalization, across the range of observed intensities. All normalization methods reduce the variability among replicates, compared to the unnormalized data. Interestingly, the complete QN appears to do a more aggressive adjustment (a greater reduction of variance), although Figure 2C shows that it does not fully normalize the control probes.

Reducing certain variability alone is only one side of the story and never enough to show the benefit of one normalization method over another. We also want to make sure that variations of interest, i.e, the actual biological variations, are preserved in the normalized data. We proceed to compare the cross tissue type variances. A good normalization would reduce technical variation among replicates, but retain real biological variation. Figure 10 shows that the complete QN reduces a lot of the between tissue variance, while the subset QN preserves the variation at the same level as the unnormalized data. This suggests that the complete QN may have paid the price of reducing signal in order to reduce noise, and may have over

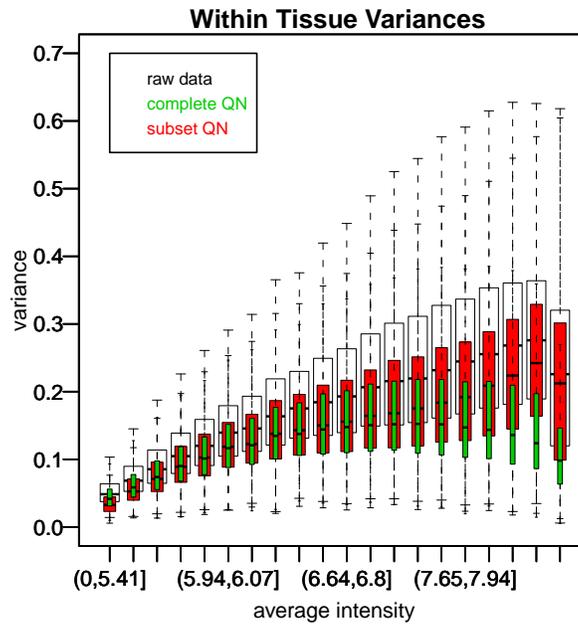


Figure 3: The within-tissue variance comparison. For each probe, the biological replicate variances for each of the 11 tissue types are averaged to give a pooled within-tissue variance. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the within-tissue variances from before and after normalization are overlaid for easy comparison.

adjusted in this example.

In order to evaluate the variance and bias trade-off of these normalization methods, we compare the between- and within-tissue variances of each probe. Since a greater extent of differential expression is expected between different tissue types than between biological replicates of the same tissue, we compute the ratio of between and within tissue variances. Because there are real biological variations even between replicates, a good normalization method may not have the greatest reduction of variance, but will increase the ratio of between and within tissue variances. In Figure 11 we compare the variance ratio over the range of average intensity. The ratio in the subset QN group is greater than that from the complete QN over the entire range of log intensities. The loess normalization result is again very similar to that from complete QN.

All the above results are done at the feature level. Since this is an experiment on the transcriptome, we also summarized the data at feature set level, using the RMA (Irizarry *et al.*, 2003) method. We compared the within-, between-tissue variances and the ratio of variances again. The results are very similar to those shown in Figures 3 - 11, and are provided in the supplementary file.

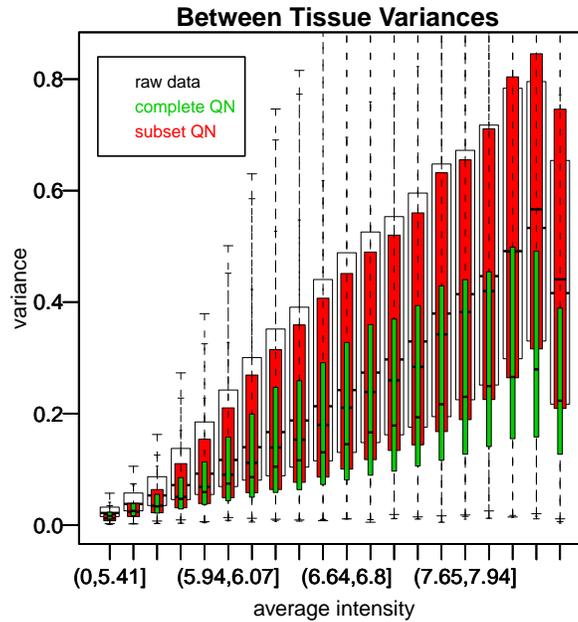


Figure 4: The between-tissue variance comparison. For each probe, the average intensity from the triplicates of each tissue type is computed and the variance of the tissue average intensities is computed. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the between-tissue variances from before and after normalization are overlaid for easy comparison.

## 5 Discussion

In this paper we present a normalization method using a subset of antigenomic features. The original purpose of including these features is to measure the extent of nonspecific binding, in order to estimate and adjust for background. However, they also serve as great source for normalization for two reasons. First, they are not expected to hybridize to the target genome or transcriptome, thus the observed intensities on these features reflect the systematic variation that we hope to remove, and not the biological signal that we hope to preserve. Second, the empirical observation shows that intensities from these features span almost the entire range of all intensities, so that we do not have to extrapolate much in the normalization. Using these probes as a basis for normalization allows us to avoid making assumptions on how the biological signal behaves in various samples. This is especially useful in situations when we are not comfortable with the usual assumptions such as the majority of genes do not have significant differential expression or symmetric up- and down-regulation. Examples include radiation experiments that disrupt the transcription of a large fraction of genes (Rea *et al.*, 2003) and asymmetric expression regulation under stress (Weber *et al.*, 2006).

The example we used in the paper is a gene expression experiment. However, this normalization method is

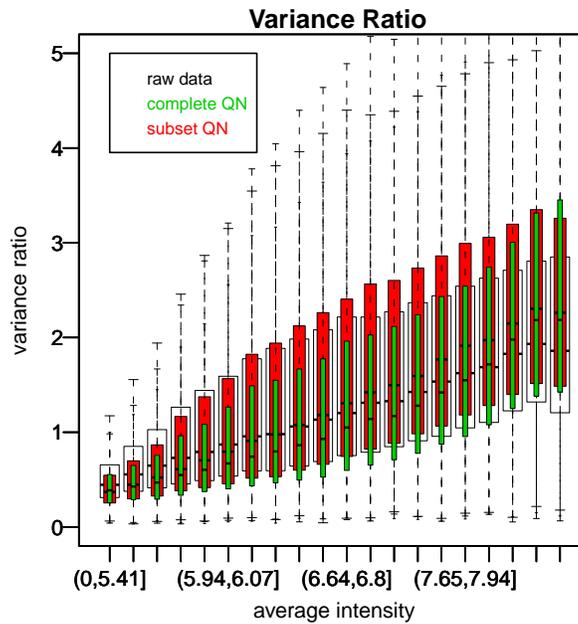


Figure 5: The ratio of between and within-tissue variances stratified by average raw intensities. Boxplots of the variance ratio from before and after normalization are overlaid for easy comparison.

applicable to other applications. For example, in studying epigenome, often a sample enriched in methylated regions is compared to a sample of genomic DNA. The methylation enriched samples tend to have different composition than the reference genomic sample, probably biased towards GC-rich regions. To assume that the overall distribution of intensities in these two types of samples to be the same, as done in complete QN, is less realistic than in many gene expression experiments. Similar problem applies to ChIP-chip experiments. Using the subset QN methods, samples with large scale differences expected can still be normalized.

Another benefit of the subset QN is that we no longer require the number of signal bearing probes to be the same across samples. This would make it easier to handle missing values, or even combining data from different platforms, as long as the same set of control features are used. This flexibility of subset QN also allows it to be applied to other high throughput technologies, as long as a subset of control features can be identified.

## References

- Amaratunga, D. and Cabrera, J. (2001). Analysis of data from viral DNA microchips. *Journal of the American Statistical Association*, **96**(456), 1161–1170.
- Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

- Fraley, C. and Raftery, A. (2009). *mclust: Model-Based Clustering / Normal Mixture Modeling*. R package version 3.2.
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**(3), 307–315.
- Irizarry, R. A., B. Hobbs, F. C., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
- Mchler, M. (2007). *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*. R package version 1.0-7.
- Rea, M., Gregg, J., Qin, Q., Phillips, M., and Rice, R. (2003). Global alteration of gene expression in human keratinocytes by inorganic arsenic. *Carcinogenesis*, **24**(4), 747.
- Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. H. R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.
- Theillin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. *Journal of biotechnology*, **75**(2-3), 291–295.
- Tseng, G., Oh, M., Rohlin, L., Liao, J., and Wong, W. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.
- Weber, C., Guigon, G., Bouchier, C., Frangeul, L., Moreira, S., Sismeiro, O., Gouyette, C., Mirelman, D., Coppee, J., and Guillén, N. (2006). Stress by Heat Shock Induces Massive Down Regulation of Genes and Allows Differential Allelic Expression of the Gal/GalNAc Lectin in *Entamoeba histolytica*? *Eukaryotic Cell*, **5**(5), 871–875.
- Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, **99**(468), 909–917.
- Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4), e15.
- Yang, Y. H. J., with contributions from Agnes Paquet, and Dudoit, S. (2007). *marray: Exploratory analysis for two-color spotted microarray data*. R package version 1.22.0.



## 6 Supplementary Figures

The following figures compare the within- and between- tissue variances, and the variance ratio from data before and after the sub QN or loess normalization. The comparison is done on feature intensity level. Compared to the sub QN, the loess normalization does a more aggressive adjustment and reduces the within tissue variance more (Figure 1), but at the price of reducing between tissue variance (Figure 2). The sub QN preserves the between tissue variance, such that the between/within tissue variance ratio is in general higher (Figure 3).

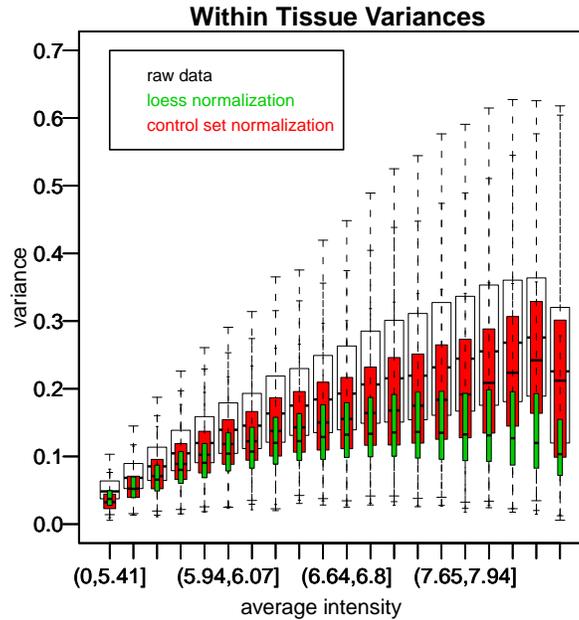


Figure 6: The within-tissue variance comparison for sub QN and loess normalization. For each probe, the biological replicate variances for each of the 11 tissue types are averaged to give a pooled within-tissue variance. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the within-tissue variances from before and after normalization are overlaid for easy comparison.

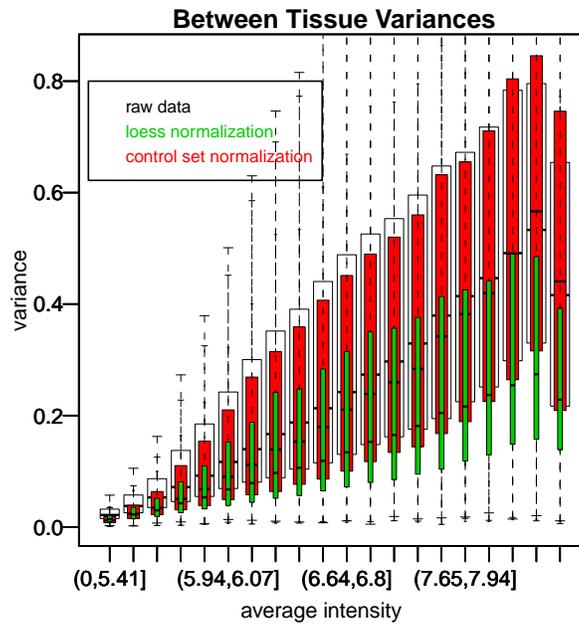


Figure 7: The between-tissue variance comparison for sub QN and loess normalization. For each probe, the average intensity from the triplicates of each tissue type is computed and the variance of the tissue average intensities is computed. The probes are stratified by average raw intensities into 20 equal-sized groups. Boxplots of the between-tissue variances from before and after normalization are overlaid for easy comparison.

The following figures compare the within- and between- tissue variances, and the variance ratio at the gene level. The sub QN reduces within tissue variance (Figure 4) and also preserves the between tissue variance (Figure 5), such that the between/within tissue variance ratio is in general higher (Figure 6).

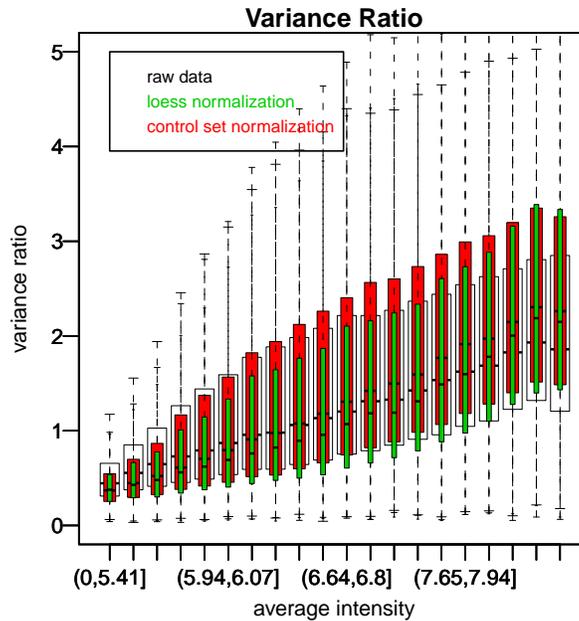


Figure 8: The ratio of between and within-tissue variances from sub QN and loess normalization. Ratios are stratified by average raw intensities. Boxplots of the variance ratio from before and after normalization are overlaid for easy comparison.

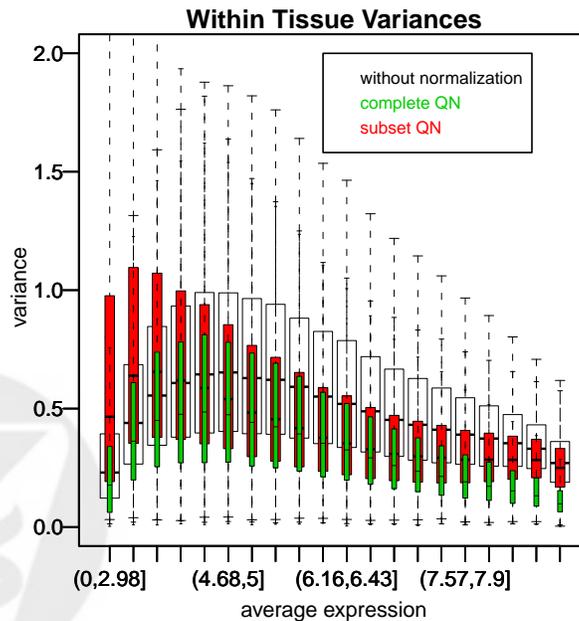


Figure 9: The within-tissue variance comparison at the gene expression level. Gene level summarization is calculated with RMA and genes are stratified by average average expression into 20 equal-sized groups. The genes are stratified by average expression from RMA at default setting.

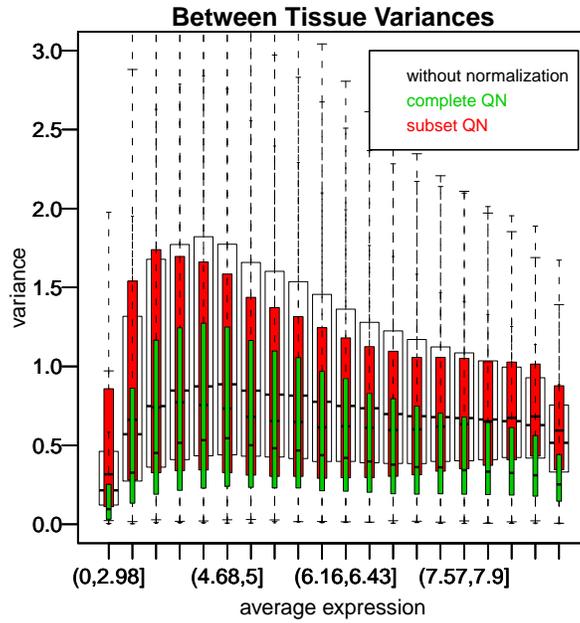


Figure 10: The between-tissue variance comparison at the gene expression level. For each gene, the average expression from the triplicates of each tissue type is computed and the variance of the tissue average expression is plotted. The genes are stratified by average expression from RMA at default setting.

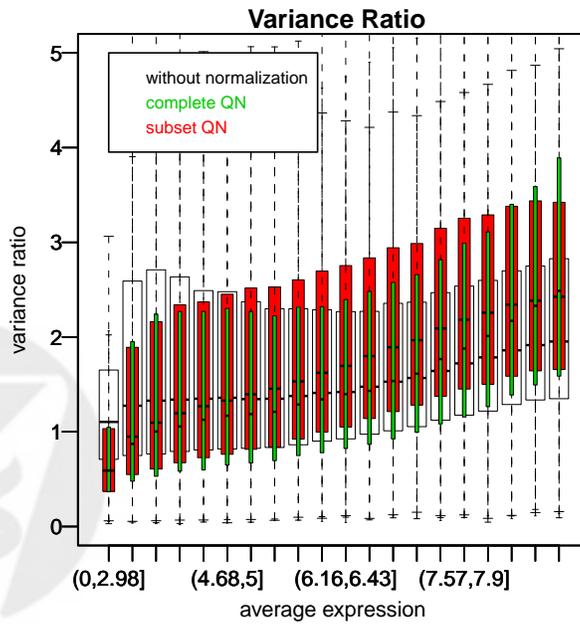


Figure 11: The ratio of between and within-tissue variances at the gene level.